AD NUMBER

AD817232

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Critical Technology; APR 1967. Other requests shall be referred to Aerospace Medical Research Labs., Wright-Patterson AFB, OH 45433.

AUTHORITY

AFHRL ltr, 18 Aug 1971

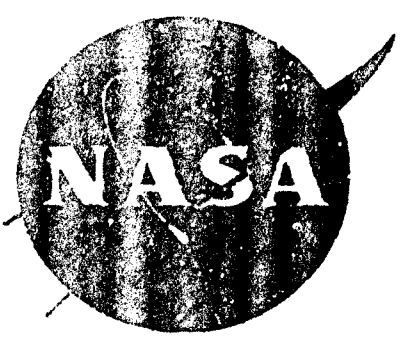
RETENTION OF SIMULATED LUNAR LANDING MISSION SKILLS: A TEST OF PILOT RELIABILITY

THEODORE E. COTTERMAN, PhD

MILTON E. WOOD

APRIL 1967

JOINT NASA/USAF STUDY



AEROSPACE MEDICAL RESEARCH LABORATORIES
AEROSPACE MEDICAL DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Requests for copies of this report should be directed to either of the addressees listed below, as applicable:

Federal Government agencies and their contractors registered with Defense Documentation Center (DDC):

DDC
Cameron Station
Alexandria, Virginia 22314

Organizations and individuals receiving reports via the Aerospace Medical Research Laboratories' automatic mailing lists should submit the addressograph plate stamp on the report envelope or refer to the code number when corresponding about change of address or cancellation.

Do not return this copy. Retain or destroy.

700 - June 1967 - CO192 - 37-910

AMRL-TR-66-222

RETENTION OF SIMULATED LUNAR LANDING MISSION SKILLS: A TEST OF PILOT RELIABILITY

THEODORE E. COTTERMAN, PhD

MILTON E. WOOD

FOREWORD

This report was prepared in direct support of planning for the Manned Orbital Laboratory and as a part of the documented effort of the Training Research Division on the problems of long-term skill retention. The in-service planning, monitoring, analyses, and writing were accomplished largely under task 171003, Human Factors in the Design of Systems for Operator Training and Evaluation. This task is a part of project 1710, Human Factors in the Design of Training Systems, under the direction of Dr. Gordon A. Eckstrand. The funds needed for the contractual data collection, preliminary analyses, and reporting accomplished by Martin-Marietta Corporation (Baltimore Division) were drawn from resources available for research on Human Performance (Budget Code 780E). Most of these funds were specially allocated for the purpose and the rest came from those assigned to project 6114, Simulation Techniques for Aerospace Crews Training, under the direction of Mr. Carl F. McNulty.

The actual test, which was in two phases, was conducted by the Martin-Marietta Corporation at its Baltimore Division under National Aeronautics and Space Administration (NASA) contracts NASw-1034, Research on Pilot Skill Retention for Manned Flight and NASw-1319, Test of Pilot Retention of Simulated Lunar Mission Skills. Dr. Milton A. Grodsky, Manager of the Manned Machine Engineering Department, arranged the details of the test, the data collection, the preliminary analyses and the preparation of preliminary reports (Grodsky et al, 1964, 1966a). He was supported in the work by several associates, especially, J. A. Mandour, D. Roberts, J. T. Warfield, and T. M. Flaherty. Although secured by Air Force funds, the supporting contracts (as indicated by their designation) were arranged by NASA on behalf of the Air Force. This was very desirable because the work complemented directly and required the use of data obtained under the NASA contracts NASw-833 and NASw-1187, Human Reliability Program, also with Martin-Marietta Corporation, Baltimore Division. Dr. Heber Moore, initiator and technical monitor of the NASA contracts on human reliability, graciously attended to arrangements for this program as well, participated in initial meetings to get the program underway, provided useful technical suggestions, and was an effective go-between throughout the course of the effort.

Within the Air Force, Col C. Lutman and later Col H. Allen, Air Force Systems Command liaison officers to NASA, provided helpful suggestions and encouragement and made the necessary arrangements with test personnel for their participation. The test personnel were 12 aerospace research pilots who had previously participated in the human reliability program. Those participating in the first test phase are Captains James H. Irwin, Lachlan MacleaRy, Albert L. Atwell, the late James S. McIntyre, Robert K. Parsons, and Russell J. Scott of the U. S. Air Force. Those participating in the second test phase are Captains Francis G. Neubeck, Thurlow H. Ralph, Charles H. Stone, and James M. Taylor of the U. S. Air Force and Lieutenants John L. Finley and Richard H. Truly of the U. S. Navy. All of the pilots responded to the test requirements very well and performed in a thoroughly professional manner, as was most necessary to the validity of the test.

Mr. Carl F. McNulty, whose initial urgings and support helped considerably to motivate the test, aided the initial planning, made the formal contractual arrangements for the first phase, and assisted in monitoring data collection in that phase. Miss Karen A. Simonsky of the University of Dayton, working under contract AF 33(615)-1824, Psychological Research on Operator Training, performed diligently and uncomplainingly by hand a great share of the calculations required for the main analysis. Mr. Robert J. Pelo, also of the University of Dayton and working under the same contract, accomplished many of the special interpretive analyses and served as a stimulating discussant of the analytic problems involved.

ABSTRACT

Four crews of three aerospace research pilots were tested on a simulated 7-day lunar landing mission at different intervals, approximating 4, 8, 9, and 13 weeks, after original training. The 6 weeks of training had culminated in the real time performance of the mission but, for the skill retention test the mission was compressed into a single 13-hour workday by omission of less significant tasks and waiting periods. Following the test one or three days of additional training on selected mission phases was given all crews.

The analysis of results focused attention on individual and crew performance at the end of training, in the skill retention test mission, and in the following retraining trials, as represented by 22 selected flight control parameters distributed over nine mission phases. By the use of novel analytic techniques the levels of performance observed were represented as reliabilities, or probabilities of success in meeting hypothetical criteria for the parameters. Also, for greater sensitivity to changes in capability, test and retraining performances were alternatively represented as probabilities of success in meeting the level of performance estimated achievable by each individual in 95% of his performances at the end of training.

On careful evaluation as to possible biases the obtained probabilities are taken to indicate (1) that lack of direct practice of critical tasks over 8 weeks or more in long duration space missions will result in unacceptable skill deterioration unless suitable remedies are sought in design and operational planning; and (2) that aerospace research pilots are capable of performing the type of mission used in this study, providing extreme care is given to their training and their individual performance reliability is demonstrated. Needs for further research on skill retention are indicated and the advantages of the novel analytic methodology used are stated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF FIGURES

RETENTION OF SIMULATED LUNAR LANDING MISSION SKILLS: A TEST OF PILOT RELIABILITY

SECTION I

INTRODUCTION

The requirement to perform specific duties and tasks after extended periods during which they are not performed is a familiar experience in both military operations and everyday living—and, incomplete recall under these circumstances is a familiar sequel. However, the need for long retention of critical knowledges and skills has seldom been so dramatically emphasized, or become such a matter of importance, as it has in relation to planning for extended space operations. For, it seems evident that in the space context there will be in the foreseeable future circumstances in which there is limited or no opportunity to engage in practice operations for training. Furthermore, there is a significant difference between routine activities and space operations that gives additional reason for interest in retention.

In everyday living and in routine military operations, there is often a considerable tolerance for error so that the quality of performance may vary widely and still be counted acceptable. But, in current space systems (as in the newest high performance weapon systems), the initial and operating investments are so large, the criticality of the missions so great, and the risks of failure so numerous and severe as to seriously call into question traditional tolerances for human error. In this context, then, of more stringent human performance requirements in increasingly lengthy missions, perhaps not inherently affording opportunities for practice of critical tasks, it is important to ask, Would operator performance be sufficiently degraded through the process of normal forgetting as to require some kind of special remedial attention?

Of course, what one is really interested in, is the more specific question, Would the forgetting of critical skills on the part of this (or these) particular operator(s) in this particular system performing this particular mission so adversely affect the probability of mission success as to warrant special attention in design? The question itself clearly implies that certain kinds of information must be available if an accurate answer is to be given. First, it is necessary to know what the relationship is between mission tasks performed by the operator and the probability of mission success. Those tasks which are most relevant to mission success and which carry stringent performance requirements are naturally of primary interest. Second, it is necessary to know the nature of these critical tasks in sufficient detail so that accurate estimates about their retention can be derived. Some kinds of tasks are more quickly forgotten than others. Finally, it is necessary to know the capabilities of the operators—how well they can acquire and retain such skills. Even very competent test pilots differ in their skills at the various kinds of activities which must be performed on a given mission and in overall capability. Furthermore, the competent individual seldom is capable of performing all aspects of a mission equally well.

If this is the kind of information needed for an accurate answer to the question of skill retention then it is appropriate to ask, How can detailed information on task criticality, specific critical tasks, and operator skill acquisition and retention be obtained? Presumably, information on the criticality and nature of tasks would be available as design data. But, to obtain information on operator skill and skill retention rather obviously seems to require a test in the actual mission. This, of course, is ruled out as impractical. What, then, are the alternatives to such an operational test?

Naturally, one thinks of simulation techniques as means of providing realistic operations for test, design, and training purposes. To obtain the needed information on pilot skill retention they might be used in either of two ways. One method would provide realism through the operation of one or more specially modified flight vehicles so that at least the performance of certain, presumably critical, mission segments could be studied. The other would provide realism through the operation of a mission simulator. But there are specific disadvantages in the use of either technique in that the mission context cannot be effectively provided with them. The sophisticated mission simulator may indeed provide the full range of mission tasks properly interlaced, shared and sequenced in time with those tasks of special interest and a mission of full duration. However, it does so at the expense of some realism in task cues and, more importantly, realistic hazards of the actual operation. On the other hand, special training vehicles presently fall far short of providing the full range of tasks organized and temporally related in a realistic fashion in missions of anything near realistic durations. Also, even the flight risks are not comparably great. Thus, the interpretation of any test results obtained by these means must compensate somehow for the lack of realism.

There are still other difficulties involved in gaining information useful for deciding about the skill retention problem if this is to be done very early in the process of system development. At an early stage the human tasks have not yet been clearly identified and defined and their criticality is not known. Furthermore, the operational personnel for specific missions generally have not been identified. As a result, it is impossible to obtain the relatively precise information needed on the proper personnel performing relevant tasks or to interpret the exact significance of any skill losses observed. And, conclusions drawn therefrom become generally weaker and less useful for decisions.

In view of such difficulties in assembling sufficient information from direct tests for early decisions, it may then be supposed that the problem can be conveniently handled on the basis of previous experimental studies of skill retention. There have been a number of experiments conducted since 1900 on this topic (Naylor and Briggs, 1960), but few useful principles for design have been or can be extracted from that research. The main reasons are that simple tasks hardly representative of operational tasks have been used, the subjects used in these studies have not been typical of individuals who will be called upon to perform space missions, and the validity of many studies is doubtful because of unsolved methodological problems in conducting them. Thus, from this available information it may be concluded merely that some decrement may occur as a function of (1) the nature of the task, (2) the degree of skill acquired, (3) the retention interval, and (4) the characteristics of the performer. Little else may be asserted.

Purpose of the Study

This study was one of several recent attempts by the Behavioral Sciences Laboratory to gain better data on the magnitude of the skill retention problem in space operations. As such, it may be placed on the middle ground somewhere between the last alternatives just outlined. For, a definite effort was made to secure more representative data by having test personnel, representative of those who will perform space missions, perform tasks like those to be encountered in space. Because of this the results are more directly applicable than most available data in the literature on the topic. On the other hand, because no real system was involved in the test, exact information on task criticality was lacking and absolute significance for mission success cannot be determined. Therefore, the results must be considered relative to the limitations of the simulation used and the prior skill levels attained by the test personnel. This may not be as serious a drawback as might be supposed. Relative information may be more useful anyhow because there is no assurance that specific tasks, task criticalities, operating environments, and skill levels will run parallel in different systems.

More specifically, the intent of the study was to obtain a better estimate than was otherwise available of the degree of skill loss in typical space vehicle operations which may be expected over periods up to 3 months duration. For this purpose a relatively complete simulated space mission was used, as performed by specially trained aerospace research pilots.

Intent and Limitation of this Report

Preliminary reports of the study (conducted in two phases) have been prepared by M. A. Grodsky, J. A. Mandour, et al (1964), and by M. A. Grodsky, D. Roberts and J. Mandour (1966a), who were directly responsible for its execution. However, a different, more intensive analysis of the data than is presented in these reports seemed advantageous by way of providing for a more precise statement of results and additional interpretations. Also, as a convenience to users of the information, a single description of the study seemed desirable. Accordingly, this report is intended as a relatively complete description of the study with emphasis upon analytic results and their implications. Some details on tasks, instrumentation, and test procedures have been omitted and for those the reader may refer to the preliminary reports and the closely related report by Grodsky, et al (1966b).

SECTION II

THE NATURE OF THE TEST

In brief, the test required the performance of a simulated lunar landing mission and, additionally, selected tasks from that mission by four research pilot crews that had been trained various periods of time earlier in the mission tasks. The simulated test mission was performed in compressed time so that intervals requiring no critical mission activities were greatly shortened. Also, each crewmember performed as crew commander in each main phase so that measures were obtained for each individual of each crew. Multiple measures of performance were obtained for each phase (with a few exceptions) and the several kinds of task parameters involved and related to the previously observed performance. In this section the details of the test concerning the test personnel and their previous training, the test performance requirements, the instrumentation, and the procedures used for testing and data analysis are briefly presented.

The Test Personnel

The test personnel consisted of 12 aerospace research pilots who were all graduates of the Aerospace Research Pilots School at Edwards Air Force Base. At the time of test 10 had the rank of Captain and two were U. S. Navy Lieutenants. They were currently assigned to experimental or instructional flying positions with the Air Force, and have bachelors degrees in either engineering, physical science or military science. Their background includes considerable military and flying experience, especially in fighter aircraft, as shown by the listing in table I.

Before testing, the 12 pilots had participated, as crews of three, in extensive training in the simulated mission. This training was accomplished as a necessary part of the National Aeronautics and Space Administration study of human reliability described in the report by Grodsky, et al (1966b). Two of the crews had trained consecutively in the late spring and summer of 1964 and the other two had trained consecutively in the summer and fall of 1965. This circumstance made it possible to obtain a natural variation in retention interval by the simple expedient of consecutively testing each pair of crews later. Thus, the first two crews returned at such a time as to provide retention intervals of 4 and 9 weeks and the second two crews returned at such a time as to provide retention intervals approximating 8 and 13 weeks. Accordingly, the crews are referred to, for convenience, as C-4, C-9, C-8 and C-13. Similarly, individual pilots (P) of the crews are designated by a number (1, 2, or 3) indicating the order in which they performed in test. Thus, P-91 is the first pilot tested in the 9-week crew and he was followed by P-92 and P-93.

TABLE I

Data on the Test Personnel

Pilot	Age (Years)	Years of Service	Flying Hours		
			Fighter	Bomber	Total
41	34	11	2800	200	3000
42	32	10	2200	400	2600
43	30	10	1760	715	2475
91	34	12	3175	150	3325
92	35	13	1600	800	2400
93	33	10	2720	60	2780
81	29	9	2900	100	3000
82	31	9	1620	0	1620
83	31	9	1600	0	1600
131	34	14	2190	1590	3780
132	27	6	1135	315	1450
133	33	10	4500	0	4500
Mean	31.9	10.3	2350.0	360.8	2710.8

The Simulated Mission and Previous Training

The simulated mission for which the crews had previously received training was a 7-day lunar landing mission. The simulation began with an ascent (automatically controlled to 100,000 feet) to a 100 NM parking orbit and continued through the following succession of major activities:

translunar insertion
transposition of flight modules
position determinations and midcourse corrections
lunar orbit insertion
lunar landing
lunar ascent
transearth insertion
position determinations and midcourse corrections
earth entry

For measurement purposes it terminated with the deployment of the drogue chute at an altitude of 25,000 feet. Interspersed appropriately among these major activities were a number of secondary checkout and preparatory activities as well as simulated emergencies.

Most of the simulated flight was accomplished by the crew of three from within a command module. Crewmembers were nominally designated commander, navigator, and engineer, but each member had his turn at performing each important activity. The mission plan was arranged so as to provide normally on-duty periods of 2 hours, off-duty periods of 2 hours, and two 4-hour sleep periods every 24 to 26 hours. Sleep periods were generally preceded and followed by off-duty periods. Exceptions to this scheduling occurred in the lunar landing, exploration, and takeoff phases, which were accomplished by two of the three crewmembers from aboard a separate excursion module. For these phases, crewmembers were on continuous duty from 8 to 14 hours, depending on position. The result was that over the complete mission each crewmember spent approximately 70 hours on-duty, 50 hours off-duty, and 50 hours sleeping.

Critical Flight Activities. The critical flight activities which had to be accomplished for a successful mission are briefly characterized as follows:

1. Translunar insertion. After about 2 hours in parking orbit, a remaining boost propulsion unit is relit to obtain the additional velocity (ΔV) required to escape earth orbit and enter a translunar trajectory. This requires interrogation of a guidance computer for information (on vehicle attitude, initiation time, and ΔV), proper attitude control, and timely initiation and cutoff of propulsion in accordance with the information provided.

2. Transposition. Beginning with an initial configuration consisting of first the command module (CM), then a service module (SM) containing flight propulsion, then the lunar excursion module (LEM), and a remaining boost propulsion unit, a rearrangement of modules is accomplished. This requires first initiating an automatic separation sequencer that separates the SM from the LEM, jettisons the LEM adaptor, and accelerates the CM-SM combination forward, and stops it. Then the pilot pitches the CM-SM 180°, flies it backward with reference to a meter display, and docks it against the LEM upper hatch. The remaining boost propulsion unit is then jettisoned.

3. Position determinations and midcourse corrections. Position determinations and any necessary midcourse corrections are made at four times during both the translunar and the transearth coast phases. These involve making optical sightings with a navigational instrument to measure the angle between an earth or moon landmark and a star, as well as star sightings to obtain data for the guidance computer. Using the automatically calculated values, the inertial measuring unit is then aligned. On the basis of the corrected navigational information, as analyzed by the computer, the pilot then makes the indicated velocity increments in the appropriate direction as to modify the trajectory toward the desired path.

4. Lunar orbit insertion. On the far side of the moon, at periynthion, retrothrust is manually applied to reduce the velocity appropriately to achieve an 80 NM circular lunar orbit. The CM-SM combination remains in this orbit while the lunar excursion is carried out.

5. Lunar landing. The lunar landing sequence is composed of several distinguishable subsequences. Operated by two crewmembers (while the third remains in the CM), the LEM is first separated from the CM, translated clear of it and stabilized in the same circular orbit. Next, at 104° central angle from the landing site with the LEM reoriented toward the center of the moon, descent propulsion is fired at the proper time to achieve an elliptical orbit with a perilune 50,000 feet over the landing site without change in orbital period. A ballistic descent trajectory is then followed with attitude being controlled to permit radar tracking of the CM. At the proper time, braking is carried out by orienting the LEM and firing the landing engine to provide a thrust vector—the crew's line of sight being 90° relative to both orientation and vector and toward the lunar surface. Pitch and roll adjustments are used to control altitude and lateral displacement and thrust level is varied to achieve zero velocity directly over the landing site. The vehicle is then pitched up to normal attitude and thrust adjusted for a hovering position at 1000-2000 feet. Finally, thrust is reduced to achieve an acceptably low rate of descent to the landing site with translational corrections made to avoid displacement from the exact touchdown site desired.

6. Lunar ascent. Upon completion of exploration and other duties on the lunar surface, the LEM crew prepares for ascent by separating the landing engine, abandoning it at the site. The ascent engine is fired with the vehicle oriented for a vertical trajectory, yaw being introduced during initial ascent to obtain downrange views. Pitch is then introduced to achieve a desired ascent profile to 50,000 feet and velocity is adjusted to achieve a Hohman transfer orbit with apilume at the CM orbital altitude. The sequence is initiated at a time which will result in the LEM being ahead of the CM in orbit. Pitch and roll adjustments are used to control altitude and out-of-plane motions. Next, with the engine off and the vehicle coasting around the moon gradually gaining altitude, the vehicle is oriented so that the CM can be acquired and tracked by radar and seen (at about 20 NM separation) by means of flashing beacon. Rendezvous is then accomplished by using translational thrust to adjust the orbital velocity vector to match that of the CM, less a closing range rate. This rate is gradually reduced for stabilization in the same orbit with a small separation distance. Finally, docking is accomplished by controlling the LEM attitude and translation to obtain closing at a very low rate with the LEM forward hatch and the CM forward hatch aligned. Upon transfer of the lunar excursion team to the CM, the LEM is left in lunar orbit.

7. Transearth orbit insertion. Sufficient additional velocity is obtained by firing the service engine to escape the lunar orbit and enter a transearth trajectory. Attitude control and timely initiation and cutoff of the engine according to data provided by the guidance computer must be accomplished. Subsequently, position determinations and necessary midcourse corrections are required.

8. Earth entry. The earth entry sequence is begun by jettisoning the SM and orienting the CM base forward before the earth's atmosphere is reached. Upon entry into the atmosphere, the trim attitude of the vehicle provides a constant lift-drag ratio of 0.5. Roll control is used as the sole means of controlling the trajectory—deflection in the direction of the lift vector being obtained in this way. At 25,000 feet the heat shield is jettisoned and a drogue chute deployed, and at 15,000 feet the main chute is deployed.

It is evident that these critical mission activities involved a variety of tasks. These may be conveniently categorized as flight control, switching and information handling (i.e., procedures), or navigational tasks. Flight control and switching tasks are involved in all of the major activities, but the navigational task is required only in the process of making position determinations. In this simulation, the flight control was almost strictly manual, in that automatic control was limited to an attitude-hold function relative to fixed inertial space or to local vertical in orbit. Changes in the attitude of either vehicle, initiation and cutoff of the engines, and translatory control were all under direct pilot control. Of course, necessary flight information was available through appropriate displays and through interrogation of the simulated flight computer. Checklists, customized by the crews for their own use, were available to and used by them.

Other activities. In addition to the major activities just described, the crews were realistically required to carry out other actions appropriate to the simulated mission and study purposes. Adjunctual, of course, to the main activities they had certain preparatory checks to run. Also, periodically malfunctions were introduced for which emergency procedures were required. Isometric exercises which had been practiced earlier were performed periodically. Only freeze-dried food was provided and each item each meal was rated. Biomedical measures of blood pressure, oral temperature, and pulse rate were regularly taken by the individual crewmembers. Feces and urine samples were packaged for later analysis. While on the lunar surface the LEM crew identified rock samples, photographed a simulated surface display and described it orally and in writing. However, such activities as these are not of direct relevance to the test of skill retention presently being described, beyond serving to further characterize the simulated mission for which the test personnel had been trained.

Previous training. Each of the test crews received 5 weeks of specialized training for the simulated mission and then, in the 6th week, performed the simulated mission in real time. Thus, for the purpose of the retention test, they may be considered to have received six weeks of training, the last week being devoted to a complete performance of the actual (simulated) mission.

During the 5 weeks of premission training, a semicustomized training plan was followed to achieve an orderly progression of skill acquisition. In general, the crews worked an 8-hour day, 5 days a week in the simulator, plus

about 2 hours per day, 6 days a week at physical conditioning. The first 3 to 5 days were devoted to lectures (concerning the mission, the vehicle systems, and the displays and controls), to study of mission and system written materials, and to crew development or refinement of their checklists. A medical examination was given and physical conditioning was started at that time.

Following the initial period, each crewmember was introduced to and allowed to practice individually each main mission task, in turn. The instructor first demonstrated the task and then observed and critiqued as necessary the crewmember's initial performances of it. Realistic mission communication procedures were used during the period.

In the next portion of the training sequence, crewmembers practiced as whole sequences of tasks, the main activities involved in the mission (e.g., the lunar landing sequences). Special practice on particular tasks was interspersed as either convenient or needed.

Finally, during the last 5 to 7 days, as this training continued, preparations for the mission in the form of briefings on checklists, food, and geology were conducted and physical status assessed. In the last two days of their premission period, C-4 and C-9 performed the mission in fast time, with the coast phases eliminated, as a whole crew. This permitted some adaptation to the living arrangements and miscellaneous mission requirements as well as further direct practice of mission tasks. C-8 and C-13 performed this fast-time mission several days earlier in their training sequence.

The simulated mission was then performed in real time over a 169-hour period, as already described.

In general, both pairs of crews received the same type of training. However, the crews trained later (C-8 and C-13) did receive substantially more actual practice in the various mission phases than did the earlier crews. This was largely the natural consequence of continuing experience in using the simulator and improvement in operating routine.

Test Performance Requirements

In the test of skill retention the crews were not, however, required to perform the full 7-day simulated mission. Instead, in the interest of minimizing costs while still providing an adequate test, they were required to perform the mission in fast time—that is, with the long translunar and transearth coast periods and the position determinations and midcourse corrections normally performed in them eliminated. In addition, the systems and log checks normally performed, as well as the transearth insertion phase itself, also were not required. Since the interest was in primary mission activities, the other activities originally included also were omitted. These omissions made it possible to complete the fast time simulated test mission within about 13 hours of a single workday, while still allowing as before, each crewmember to perform each major activity. In this way, considerable test performance data were gathered with minimal simulator operating time and cost. The planned test mission schedules are reproduced in abbreviated form in tables II and III to illustrate the testing sequence. From these it may be noted that except for transposition by C-4 and C-9 pilots within crews performed the test in an invariant order.

TABLE XI

Schedule for G-4 and G-9 Test Mission

Phase	Time (EST)	Crewmember Position		
		Pilot	Nav.	Engr.
Briefing	0800	- Pilots' Office -		
Pilot Insertion and Prelaunch Check	0830	1	2	3
Earth Ascent	0900	1	2	3
Translunar Insertion	0930	1	2	3
Translunar Insertion	0950	2	3	1
Translunar Insertion	1010	3	1	2
Transposition	1030	3	1	2
Transposition	1050	1	2	3
Transposition	1120	2	3	1
LEM Status Check	1145	1 (3*)		
Lunar Orbit Insertion	1230	1	2	3
Lunar Orbit Insertion	1250	2	3	1
Lunar Orbit Insertion	1310	3	1	2
Lunar Landing and Ascent - Docking with CM	1330	3 (1*)	(2*)	
Lunar Landing and Ascent - Docking with CM	1530	1 (2*)		
Lunar Landing and Ascent - Docking with CM	1730	2 (3*)	(1*)	
Earth Entry	1930	1	2	3
Earth Entry	2000	2	3	1
Earth Entry	2030	3	1	2
Completion / Debriefing	2100	- Pilots' Office -		

* Positioned in Lunar Excursion Module

TABLE III

Schedule for C-8 and C-13 Test Mission

Phase	Time (EDST)	Crewmember Position Pilot	Nav.	Engr.
Briefing	0730	- Pilots' Office -		
Earth Ascent	0800	1	2	3
Translunar Insertion	0915	1	2	3
Transposition	0935	1	2	3
LEM Status Check	0955	1 *		3
Lunar Orbit Insertion	1035	1	2	3
Lunar Landing	1055	1 *		3 *
Lunar Ascent, Rendesvous, Docking	1155	1 *		3 *
Earth Entry	1255	1	2	3
Translunar Insertion	1340	2	3	1
Transposition	1400	2	3	1
Lunar Orbit Insertion	1420	2	3	1
Lunar Landing	1440	2 *		1
Lunar Ascent, Rendesvous, Docking	1540	2 *		1
Earth Entry	1640	2	3	1
Translunar Insertion	1725	3	1	2
Transposition	1745	3	1	2
Lunar Orbit Insertion	1805	3	1	2
Lunar Landing	1825	3 *		2
Lunar Ascent, Rendesvous, Docking	1925	3 *		2
Earth Entry	2040	3	1	2
Completion / Debriefing	2115	- Pilots' Office -		

* Positioned in Lunar Excursion Module

fact, those of C-4 and C-9 all performed a given mission segment before going on to the next segment, whereas those of C-8 and C-13 performed the entire mission in sequence before changing position. Actually, because of their effective and expeditious performance, the crews completed their missions somewhat earlier than is shown.

On the day (or days) following that in which the fast-time test mission was performed, each crew was then further required to perform, repeatedly, selected primary mission activities of special interest. As in the test mission, each crewmember performed on each of the phases so that data were obtained from all the participants. The purpose of these trials was to obtain a basis for estimating how rapidly any skill loss occurring over the retention period might be compensated for by renewed practice. Also, the additional training afforded in this way permitted a check on whether the crews had attained maximum efficiency at the end of original training (particularly those having 3 days of additional trials.

For two of the crews, receiving 1 day of repeated trials, the primary mission phases selected were transposition, LEM braking and hover, LEM docking, transearth insertion, and earth entry. For the other two crews, receiving 3 days of repeated trials, the lunar orbit insertion was substituted for the transearth insertion and the LEM separation and deorbit phase was additionally required. The number of trials given accordingly are shown in table IV, for the two pairs of crews, respectively. Because of their speedy performance, the crews were able to perform more trials than had been planned originally.

TABLE IV

Repeated Trials on Mission Phases

Phase	Trials/Crews C-4 and C-9	C-8 and C-13 *
Transposition (TEM)	2	12
Lunar Orbit Insertion (LOI)		12
Separation and Deorbit (SDO)		12 (16 for P-83)
Braking and Hover (BH)	8	16 (28 for P-131; 24 for P-132, P-133)
Docking (Dok)	6	8 (12 for P-132, P-82, P-83)
Transearth Insertion (TEI)	2	
Earth Entry (EE)	5	12

* Except for P-81 who performed only 8 SDO, 12 BH, and 8 EE phases

Instrumentation

The facilities and equipment used in this test were those used in the previously mentioned contract research on human reliability, in which the training and real time mission were performed. The major components employed were a simulated command module with associated outside displays; a simulated lunar excursion module crew compartment with associated outside displays; extensive analog computing and recording equipment, a simulation control room with panels and consoles for monitoring displays, making inputs to them and recording system outputs; and supporting office and maintenance areas. A main simulation area (about 1500 sq. ft.) contained the two vehicle simulations, the control room was adjoining, and most of the analog computing equipment was in a separate building over 1000 feet away.

Command Module. The simulated CM was a truncated 60° cone of aluminum skin on stringer and frame construction having a base of 166 inches diameter and an enclosed volume of about 400 cubic feet. It was oriented with the small end forward and the axis of symmetry horizontal on a vibration-isolated and sound-damped cradle. A hatch in the base (or rear) was normally used for entry and a hatch in the small end allowed crew transfer to the LEM, when attached to the CM. Figure 1 depicts the vehicle as seen from the right rear and shows the normal entry hatch with navigational equipment nearby. The front of the vehicle may be seen in the background of figure 2.

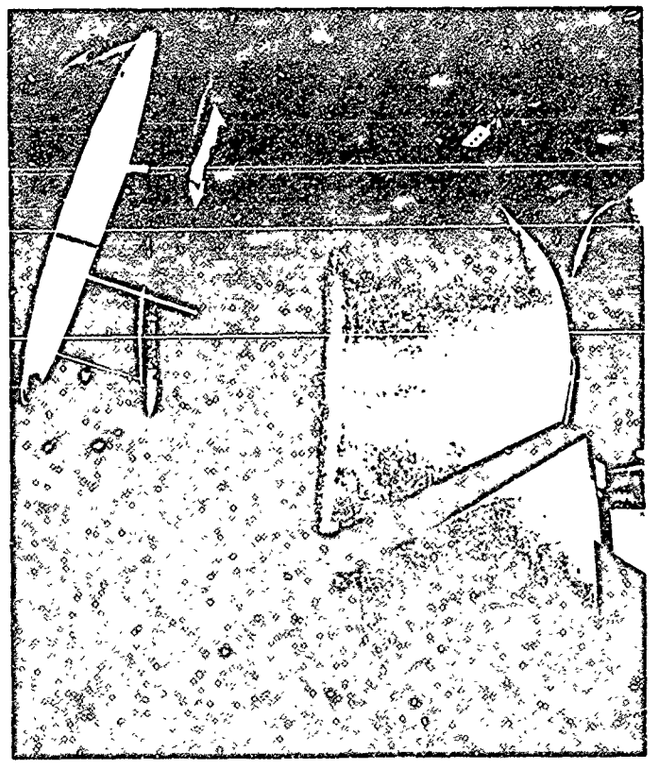


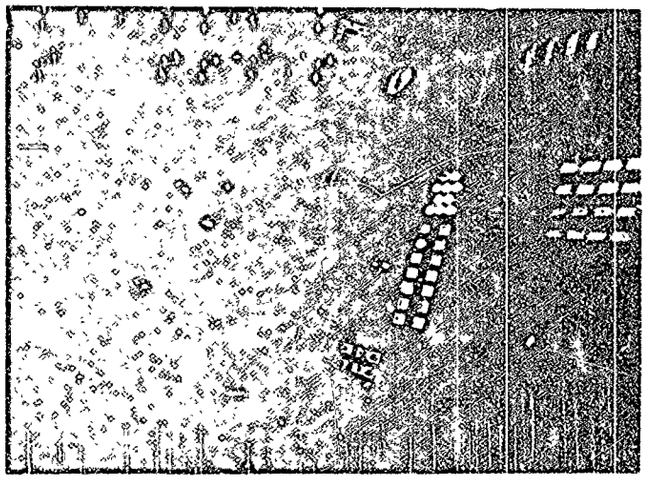
Figure 1. Right-Rear View of Command Module

The interior space was arranged by means of suitable part-floors and fixtures into several areas. The duty area was located in the upper forward portion. It consisted of a main control panel stretching completely across the vehicle, small side panels, and three seats mounted side-by-side facing the panel. The middle seat could be moved forward or backward or removed completely for use in the off-duty area. The seats were allocated to the commander, navigator, and engineer in left to right placement. In the upper rear portion was a sanitation area for waste disposal at the left and a food and water area at the right. In the lower portion were a sleeping area to the left (containing a mattress, privacy curtains, and separate lights) and an off-duty area to the right of a central bay. A navigator's station was just inside the entrance hatch in the rear bulkhead. Three TV cameras provided separate monitor views of the three crewmember stations. There were live microphones in both the duty and off-duty areas as well as individual headsets. Main rocket engine noise was simulated through a separate speaker above the duty area. The interior of the vehicle was lighted and air conditioned by integral units.

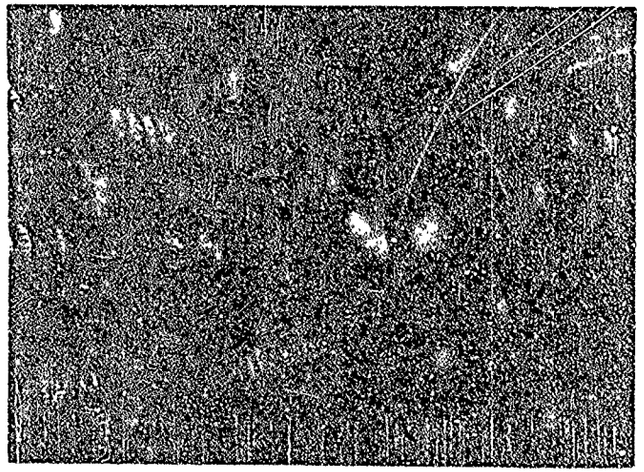
At the pilot's station (figure 3), in addition to appropriate function switches and status indicators, were the main instruments and controls for controlling the vehicle. A sidestick for translation was located on the left armrest and another for attitude control was located on the right armrest. Deflection of the translation control resulted in proportional translatory accelerations. Motions in X- and Y-dimensions produced compatible forward or sideward accelerations, respectively, and clockwise twisting resulted in downward accelerations. Deflection of the attitude control produced proportional angular velocities in the same directions as a conventional aircraft stick and the additional twist motion produced yaw in the direction of twist. A 3-axis ball indicator with pitch and yaw error needles provided primary attitude information. A digital indicator was provided for setting in desired velocity increments, as in orbital insertions and corrections. This would pulse down to zero as the thrust was introduced, at which time the pilot was to cut off the engine. For transposition a 3-needle display indicated the information on CM position relative to the LEM necessary for the pilot to translate back to it, by compensating for the indicated errors. For reentry, a 2-needle roll meter indicated the roll program and actual roll superposition of the two needles being the continuously desired status. Actually, since this status was virtually impossible to achieve, deliberate uncommanded roll corrections were required.

At the navigator's station (figure 4) were mounted a single line of sight scanning telescope, a dual line of sight sextant, a small 2-axis stick (for proportional control over shaft and trunnion angular rates of the telescope and one sextant line of sight relative to the other sextant line of sight), a small 3-axis stick for "bang-bang" control of attitude drift, and associated switches. Outside the module, operated in conjunction with the navigator's equipment was a gimbal-mounted assembly of a starsphere and slide projectors. With these, a starfield (110° visible) and one of up to six different earth or moon views were presented on a 7-1/2-foot spherical screen as a display for navigational fixes and guidance system alignments in cislunar space. The navigator's attitude drift control moved the entire scene in yaw, pitch or roll at very slow rates through the gimbal drive mechanisms.

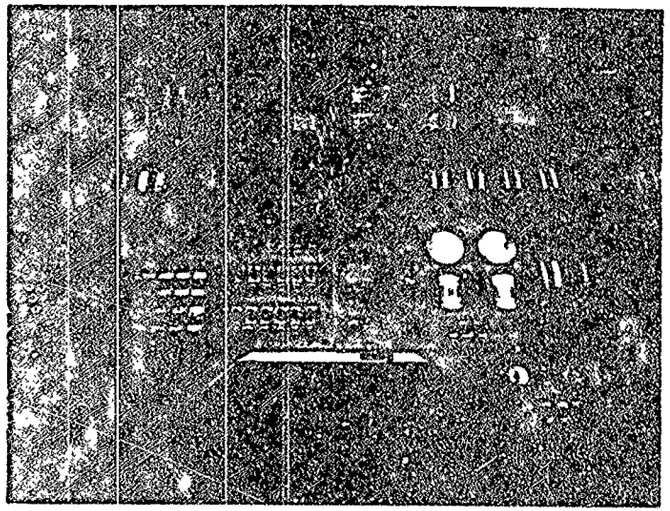
re 2. Front View of Simulation Equipment



PILOT'S POSITION



ENGINEER'S POSITION



NAVIGATOR'S POSITION

Figure 3. Main Command Module Panels

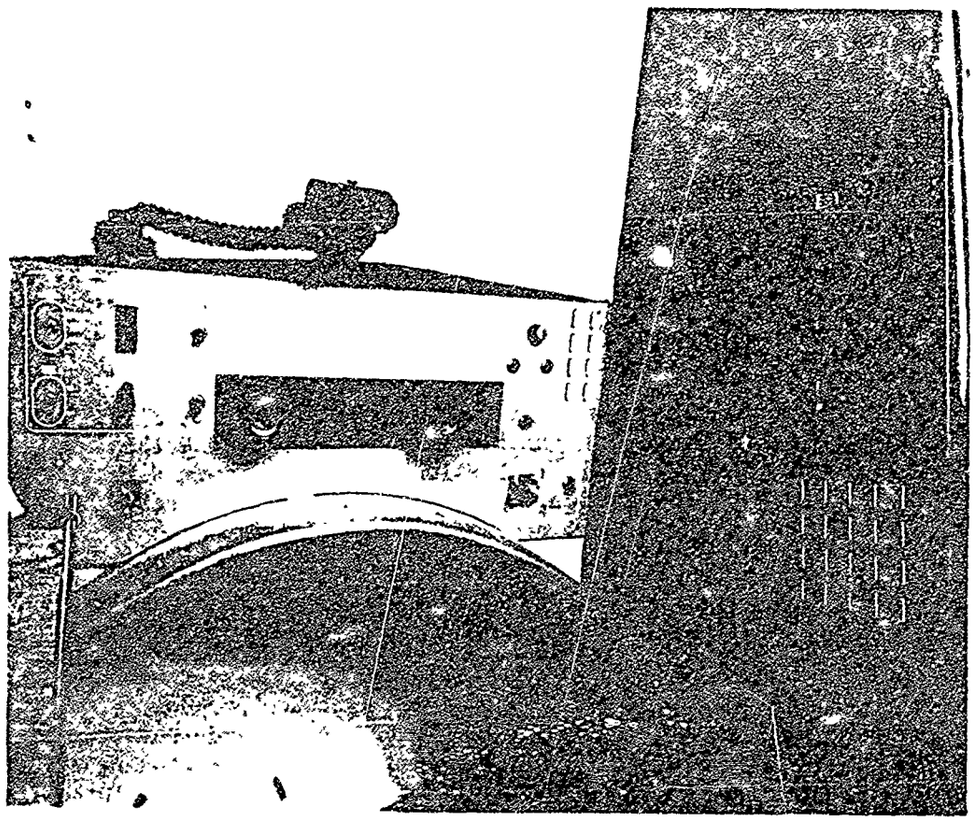


Figure 4. Navigator's Station

The entire CM simulation was programmed on one Pace 231R Computer using several economization techniques. Because of the communalities in equations for orbital insertions, midcourse corrections, transposition and earth entry, it was possible to program the rotational equations and change of velocity calculation once for all these phases. With this arrangement and additional programmed calculations, as required (for earth entry, in particular), the change from one phase to another was accomplished merely by changing certain potentiometer settings (e.g., mass, mass flow rate, thrust, moment of inertia, control gain). No more than 12 changes were required in any circumstance. Earth ascent and position determination required unique programs, but relatively little computational equipment by comparison. Then, additionally, to obtain desired information on pilot performance from flight parameters and manipulations of them, the program was separated into two patch panels. One contained all phases to earth entry and the other contained only the earth entry phase. These were arranged in such a way that different potentiometers were used, thus enabling a switch from one to the other without changes in potentiometer settings. Appropriate checkout and calibration techniques were used following changeover from one phase to another to insure a correct simulation. The computer controls were wired so that appropriate portions of the program automatically became active whenever the pilot selected a given control condition, and terminal values at the completion of the task were automatically held for readout.

LEM Crew Compartment. The simulated LEM crew compartment of aluminum skin on stringers and frames provided sufficient space for side-by-side seating of two crew members facing forward in semierect posture. The seats consisted of harnesses attached to vertical cradles of tubing. A forward hatch permitted access from and return to the CM when the LEM was attached. The compartment was supported at the center of gravity by a 3-axis, hydraulically driven gimbal system that permitted attitude changes of $\pm$ 40° in pitch and roll and $\pm$ 180° in yaw in response to computer signals. (Actually, pitch and roll were used only in docking to the limits of $\pm$ 15°.) The gimbal system was supported by a steel stand of such height that the hinge-point was 8-1/2 feet above floor level. The front or hatch end of the LEM crew compartment may be seen in the middleground of figure 2, which shows the compartment just ahead of the simulated CM, which is at the rear.

At the pilot's station the primary controls provided for flight control were an attitude stick, a translation stick, and a main engine throttle. The attitude stick was floor-mounted. Deflection of it in a given direction produced proportional angular rates in pitch, roll, and yaw. Rearward deflection resulted in pitch-up, movement to the right produced yaw in the same direction, and counterclockwise twisting produced right roll. To operate the stick it was necessary to depress a trigger which, when released, returned the system to an automatic attitude hold condition. The translation control was a T-stick projecting horizontally from the instrument panel. In one mode used for deorbit, rendezvous, and docking, deflections produced proportional translatory accelerations. In the other mode, used for hovering, deflection produced proportional translatory rates. Upward-downward, right-left, and forward-rearward motions which it permitted all resulted in translations in the corresponding directions. A trigger on the translation stick coupled the throttle to it and locked the simulated vehicle in a hover while still enabling translations.

The primary displays included a ΔV counter, a combined crossrange and crossrange-rate meter, a combined attitude and attitude rate meter, a range meter, a range rate meter, and a CRT display of downrange and crossrange displacement from the landing site. As in the simulated CM, to add velocity the pilot set the computed value on a digital meter, fired the engine at the proper time and cut off the engine when the indicated value reached zero. The range and range rate meters were interpreted with the aid of a table attached to the panel which showed the schedules of range rate with range for various throttle settings producing a vehicle stop at the landing site. The attitude control display was a large meter indicating attitude by position of a long horizontal needle and attitude rate by position of a short vertical needle—the optimum continued intersect of the two for various braking attitudes being shown by appropriate curves. Similarly, an ascent profile on the same meter guided the powered ascent. In addition to the CRT display of downrange and crossrange displacement, a downrange rate needle was also included on the crossrange meter to facilitate the hovering and landing tasks.

A spherical screen, two projector assemblies, and a translator were used to provide realistic outside views in the various phases of the lunar excursion. Separate triangular windows in the crew compartment allowed each crewmember a viewfield $\pm$ 90° in azimuth and 18 to -90° in elevation. The appropriate views were simulated by projections on a 24-foot diameter screen. A star-horizon projector mounted on a 3-axis gimbal system on the ceiling of the simulator room and driven by computer signals gave realistic indications of pitch, roll, yaw,

central angle and altitude. The starsphere of 18-inch diameter included the 2300 brightest stars. A small incandescent lamp and suitable masks mounted outside it gave the illusion of the moon and earth and moon horizons. For rendezvous, a separate projector mounted on a 2-axis gimbal support and oriented by computer signals derived from those representing the orbits of the vehicles (LEM and CM) gave a realistic view of the CM outlined by its flashing beacon which became a steady outline increasing in size with diminishing range. Coincidentally, as range approached 12 feet, through nonapparent 3/32-inch perforations (spaced 1/2-inch) in the screen, the crew viewed an illuminated translator consisting of a lightweight shell of the same shape as the CM. Then, for docking, the screen was rapidly separated along a vertical centerline and the two halves rotated backward and to the side by pneumatic pistons and hydraulic actuators. With the LEM compartment gimbals permitting pitch, roll, and yaw motions and the translator mounting permitting vertical and lateral travels up to ± 5 feet and fore-aft travel up to 12 feet, physical docking was then accomplished. Translations, by means of direct current electric motor servosystems and cable drives, were signaled by the computer on the basis of pilot deflections of the LEM translation stick. The CM was assumed stabilized in relative position. The simulated LEM compartment, the screen (partly open), and the translator (CM) may be seen in the center to rearground of figure 5. Just the sidewall of the simulated CM shown in the foreground of this view.

The LEM simulation required the use of two Pace 231R Computers and considerable associated switching and relay circuitry. Because of dissimilarities in parameter ranges, command inputs, etc., the mission was divided into a series of individual problems or phases. Also, because of the nature of the operation, two axes of references are required: an orbital, with translating origin and nonrotating axis for deorbit and rendezvous; and suborbital, with nontranslating origin for braking, landing and ascent. The equations of motion were programmed only once and stepper switches were used to go from one phase to the next. Except for the inclusion of attitude hold circuitry and rotation of the roll axis to 90°, the rotational equations were performed in the conventional way. In programming the translation equations, because of the display range requirements, careful attention was given to the problem of scale changes. These were accomplished by switching the displacements at the rescaling point of the range display and by alternately using parallel integrators and storing the variable values at a scale factor 10 times higher than the one being used. The long coasting descent and ascent phases were function-generated rather than computed. As with the CM simulation, the computer controls were wired so that terminal values at the completion of the phases were automatically held for readout.

Simulation Control and Recording. The entire simulation was monitored and coordinated from an adjacent control room, a view of which is shown in figure 6. Included were a communicator's console, a flight director's console, systems operation consoles, and data recording equipment. An intercom system connected the control room with the simulated vehicles and computer personnel. It provided for monitoring of all communication by control and computer personnel, tape recording of all vehicle-control room communications during missions, and separate communications with the observer-operator in the main simulation room. Vehicle-control communications could be delayed as a function of range. A direct telephone line was available between the flight director and computer personnel. The communicator had available the necessary communication equipment, a mission time indicator, duplicates of all CM and LEM caution and warning indicators, and a TV monitor (switchable to any of the three CM crew position

View of S[illegible]ation Equipment

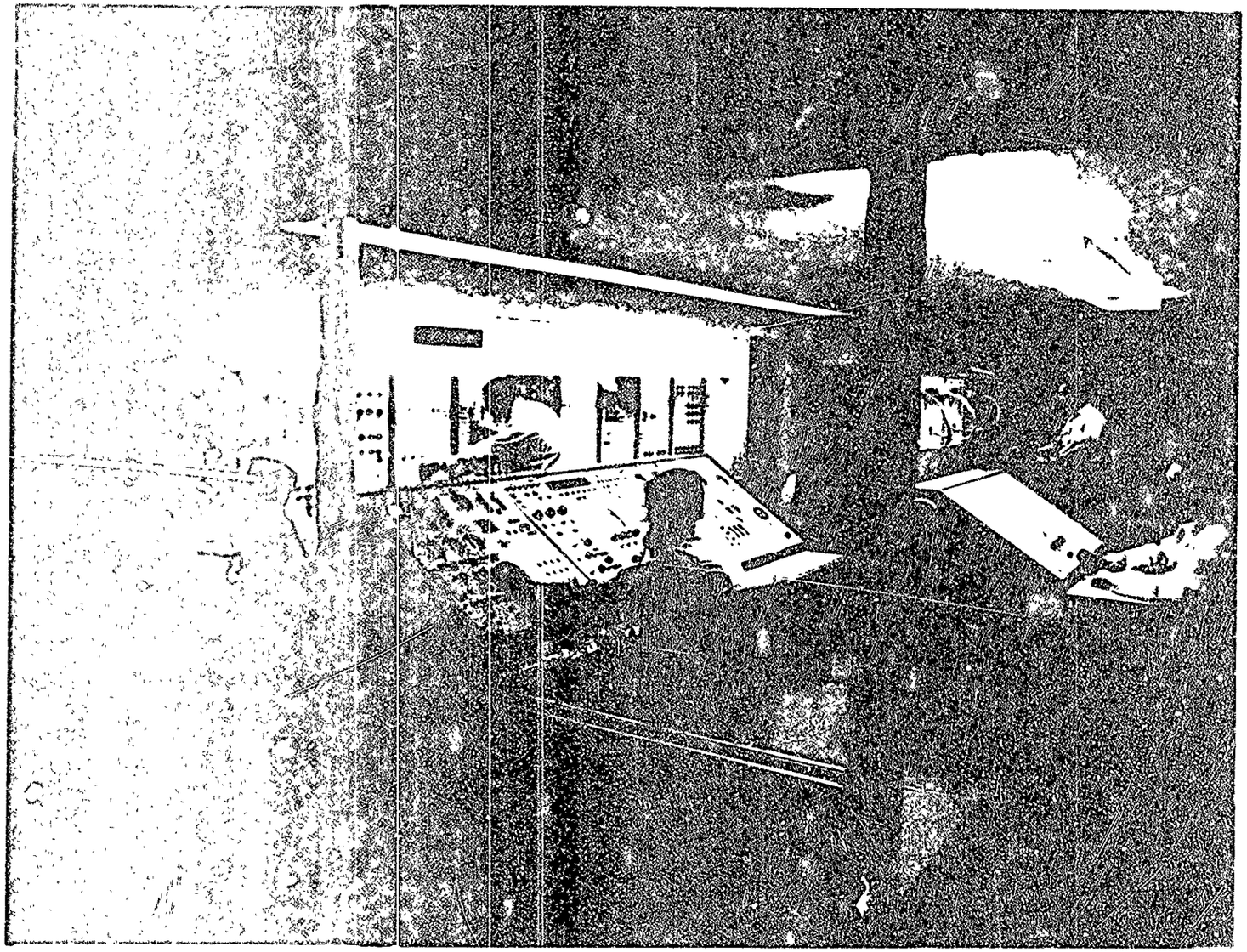


Figure 6. Simulation Control Room

monitors or a camera viewing the LEM docking. The flight director had available a similar TV monitor; duplicates of most CM and LEM flight instruments; duplicates of all CM and LEM warning and caution indicators; indicators and switches relating to control, navigation and radar systems; and controls for the mission timer and outside projection displays. Numerous systems operations consoles provided indicator lights, switches, and potentiometers corresponding to lights, switches, and meters on the simulated vehicle panels. These allowed full monitorship of the flight crew's operation and appropriate responses to it as well as the introduction of malfunction indications.

Within the control room, three 100-channel Brush binary recorders were connected to either the simulated CM or LEM compartment to record running information on switch positions and status indicator lights. A fourth was connected to the CM inflight test system. Nonflight system meter readings, motions of the navigation optics, and forces exerted during the isometric exercises were recorded with three 50-channel Consolidated Electrodynamics oscillographs. In the computer laboratory, flight measures were recorded on five Mark 200 8-channel rectilinear recorders and two 11 x 17-inch X-Y plotters and digital printouts made of relevant terminal conditions on the computers.

Limitations in the Simulation. In view of the emphasis upon assessing task skill, as such, the simulation was considered to be reasonably adequate for the purpose of this study. This may be evident from the brief description just given of the instrumentation, although the reader may wish to check on certain details in the report by Grodsky, Mandour, et al (1966b) or others to which that report makes reference. However, certain rather obvious deviations from a full, high-fidelity simulation should be noted for the record.

First, cues and conditions deriving from motion were simulated in a limited way. Although attitude and translatory motions could be made up to reasonable limits for the purpose of specific maneuvers, there was no simulation of ascent and reentry accelerations and vibrations or of impact G on landing. The condition of weightlessness also was not simulated. Second, in the life support area the environment deviated significantly from the real one in that appropriate variations in atmospheric composition and pressure, toxic gases, and radiation were not included. Also, a fully realistic handling of nutrition and sanitation was not required nor was the use of personal protective equipment required. Third, with respect to specific tasks the lunar landing was made with reference solely to instruments rather than with at least part use of an out-the-window view. Simulation of the out-the-window view was considered unnecessarily expensive in view of the study purposes. Finally, the context of mission actuality was necessarily missing with the likely result that crewmember motivations and emotional responses (particularly anxiety) were somewhat different than may be expected on a real mission.

Just what effect these departures from realism may have had on the outcome of the study is difficult to guess. There are reasonable bases for arguing that specific task performance may have been selectively improved, or selectively degraded, or even unchanged by these circumstances; and specific cases may be made and supported by previous studies of these variables. However, it is doubtful whether overall performance on a real mission of this nature would exceed that displayed in the test. The principal advantage accruing in the real mission is considered to be the resulting motivation to perform well and without failure. All the test personnel gave evidence of such an intent. Certainly,

their response to the simulation was generally favorable—the principal exception being the criticism by C-4 and C-9 of the unconventional fly-from arrangement of the landing site displacement indicator in the simulated LEM crew compartment which caused confusion in direction of control motions. Consequently, the directional response of this indicator was reversed, so as to be conventional, for C-8 and C-13.

Test Procedures

The details for arrangement and conduct of retention testing were simple and straightforward. Having previously been contacted informally concerning their availability and warned of the probable requirement for their presence, the test personnel were subsequently directed by message from Air Force Systems Command to be present for the specified period—3 days for C-4 and C-9, 5 days for C-8 and C-13.

Accordingly, and in keeping with informal word passed to them, each crew arrived at the test site for duty just after noon their first day. As scheduled in the first phase, the first crew to be tested (C-4) arrived on a Monday and concluded test duties Wednesday evening; the second crew (C-9) arrived on the same Wednesday and concluded test duties Friday evening. All testing in this phase was conducted during the 5 days of a single workweek. As scheduled in the second phase, the first crew to be tested (C-13) arrived on Monday and completed work the following Friday and the second crew (C-8) arrived the succeeding Monday and completed work the succeeding Friday. There was one exception to this general arrangement in that P-81 of C-8 was obliged, because of unforeseen other duties, to leave at noon Wednesday rather than to complete the workweek.

Upon arrival at the test site for duty, each crew was given a briefing by Air Force and Martin-Marietta personnel about the purpose, general arrangements for, and manner of conducting the test. Following this, each crew was allowed to choose how it would review the systems and tasks in the balance of time available that afternoon. Their checklists were available to them at that time and subsequently. After hearing a short review briefing on systems tasks and descriptions of any change in the simulation (these were trivial), C-4 spent about 1/2 hour, C-9 spent about 2 hours and C-8 and C-13 each spent about 1 hour reviewing their tasks and procedures. However, the crews were not permitted to view the actual displays and controls or enter the simulators. This completed the first day's test activities and crewmembers were then free to pursue their own concerns until the following morning.

On the second day, each crew then performed the simulated fast time mission discussed earlier and outlined in tables II and III. On completion each was once again free of duties until the following day. On the third day, and the following two days for C-8 and C-13, each crew then repeatedly performed selected mission phases, as shown in table IV. On completion of these at the end of the workday, they were given a very short briefing on general outcome of the test, encouraged to make any comments they chose concerning the test and their test performance, and then released to return home.

Throughout the simulated mission and mission phase performances, appropriate control of the simulation was maintained by control room personnel and realistic communications procedures were used. Certain other Martin-Marietta and Air Force personnel also were present periodically to monitor the conduct of the test, but they did not interact with the pilots during actual test performance.

Performance Measures

Having compressed the normal 7-day mission into a single workday by omitting the miscellaneous activities normally required and the long coast phases, including the navigational tasks of those phases, the measurable aspects of performance were considerably reduced. Nevertheless, there were a number of measures still obtainable, such as the various system parameters reflecting upon flight control, the associated switching requirements, and the procedural checks preceding and following various phases. However, in the retest performance of C-8 and C-13 on the days following the test mission even the procedural and switching activities were eliminated in order to obtain additional data on the apparently more sensitive flight control capability. The net result of these constraining decisions was to limit the primary utility of the data to the area of flight control. Furthermore, as described by Grodsky et al (1966a) and Grodsky et al (1966b), the data on switching and procedures do not show evident sensitivity to the retention variable. Also, being of a frequency nature and subject to various fractionations as to criticality, phase, position and so forth, it seems unlikely that valid inferences could be based upon the few cases in each properly distinguished subset. On the conclusion, then, that these data contribute little to the understanding of skill retention in this inst . they are not treated in this report, but they are summarized for the curious der in appendix I.

With respect to flight control, of course, a large number of measurable parameters were recorded, and some limitation as to which would be considered was necessary to keep the analysis and interpretation tasks manageable. Thus, having enjoyed the opportunity to comment concerning the various measurement possibilities before they were selected, it seemed appropriate to use in this analysis of retention approximately the same measures agreed to by Martin-Marietta and NASA for use in the closely related human reliability program (Contract NASw-1187). This group of measures had been carefully selected from the many possible with the intent of providing a sufficiency of information for adequate representation of the flight control performance and their general significance can hardly be questioned. The use of essentially the same measures also allows for more direct cross-comparison of findings in the two studies and of results obtained via different analytic methods. Accordingly, flight control and operations were represented by the various measures, as briefly described in the following paragraphs. The associated hypothetical system criterion for each, arrived at in the same way and in view of early data, is also included.

Translunar, Lunar Orbit, and Transearth Insertions (TLI, LOI, & TEI). In these three insertion phases the aspects of performance considered most critical are the velocity cut-off and the adequacy with which the pitch program is tracked. Any error in initial orientation of the vehicle (CM) would be promptly nulled out as a pitch error. engine-ignition may occur within a time span of 30 seconds without difficulty, and coasting attitude is not critical.

Since a specific velocity was required at cutoff, as indicated by a counter pulsing to zero, it was appropriate to simply record in feet per second the deviation from the required value. Similarly, since pitch control adequacy was reflected in minimal deviation from an optimal program, the differential was integrated over time to indicate overall accuracy. However, accuracy can be reflected in two obvious ways—as average or arithmetic mean deviation and as a variation or standard deviation about that mean. Hence, both the mean pitch error ($\bar{X}$) and the standard deviation of pitch error (s) were recorded. In the analysis presented in the following these have been combined on a trial-by-trial basis to provide an estimate of the maximal error occurring 95% of the time ($\dot{X}_{.95}$), assuming the error distribution normal. The actual relationship used is

$$\dot{X}_{.95} = \bar{X} + 1.6449s$$

In this way pitch control accuracy was represented by a single measure. The criteria selected were $\pm$ 10 fps for velocity and $\pm$ 0.1° for both pitch error mean and standard deviation, or $\pm$ 0.2645° for the combined pitch error.

Transposition (TPN). Performance of the transposition phase was considered best reflected in the displacement and displacement rate of the two vehicles at docking, the closing rate at impact of the two vehicles and the fuel required for accomplishment of the phase. Initial separation and stabilization were accomplished automatically and any error in inversion to 180° could easily be corrected. Hence, displacement (in feet), displacement and impact rates (in fps), and fuel (in slugs) were recorded. However, in the following analysis displacement was omitted from consideration on the argument that if the phase was accomplished at all, the amount of displacement would be acceptable. The corresponding criteria selected were 1 fps for both rate measures and 10 slugs of fuel.

Separation and Deorbit (SDO). Separation of the LEM from the CM and subsequent deorbit in the LEM constitute the first main phase in the lunar excursion sequence. The most critical aspects seem to be the proper orientation at the time of ignition and timely engine cutoff upon achieving the proper change in velocity. Accordingly, the measures selected as of special interest were velocity cutoff error (in fps), pitch angle error (in degrees) and yaw angle error (in degrees). The corresponding criteria were set at 2 fps and $\pm$ 2° and $\pm$ 1°, respectively.

Brake and Hover (BH). After the deorbit and subsequent coast descent (in which no piloting is required) the braking and hover phase of the lunar landing must be accomplished. This is naturally a critical phase and, at least in this simulation, a complex one having a number of measurable aspects. As a result, a variety of measures were selected for special attention. These included the percentage of available fuel consumed (an efficiency measure), the displacement (in feet) from intended touchdown site, the vertical translation or impact rate (in fps), the lateral translation or displacement rate (in fps), the roll and pitch angles (in degrees) and the attitude change rates (in degrees per second) in all axes at touchdown. However, in the following analysis several of these measures were omitted as less important for close scrutiny and the performance is represented simply by fuel consumption, displacement, and displacement and impact rates. The corresponding criteria are 95%, 200 ft, and 10 fps for both rates.

Lunar Ascent (LA). The powered ascent phase is relatively simple. Engine ignition and capsule orientation to ascent altitude present no problem. Tracking the altitude profile by varying pitch angle was not particularly difficult because of the slow rate of change involved nor was subsequent reorientation at 1,000 fps velocity. However, proper cutoff of the engine at the desired velocity is critical and, under these circumstances, apparently subject to error resulting from attention to other aspects of flight control. Therefore, velocity cutoff error (in fps) was taken to represent performance adequacy in this phase and the criterion was set at 10 fps. The following coast ascent required only the easy task of maintaining radar lock on the CM.

Rendezvous (Rend). Upon reaching an approximation of the CM orbital altitude, rendezvous with it was accomplished by acquiring a visual fix on its beacon and stabilizing in a coorbiting position relative to it by operating the LEM translation stick. This performance was considered best represented by the efficiency measure of percentage of available fuel consumed in the process, and the criterion was again set at 95%.

Docking (Dok). Like the earlier transposition phase, the subsequent docking or physical join-up of the LEM with the CM is reflected in the critical performance aspects of displacement and displacement rate and in impact rate. These were duly considered in the analysis described in the following, the criteria being 1 ft, 0.5 fps, and 0.1 fps, respectively. (Displacement was retained in this case, partly because no independent measure of efficiency, as fuel consumption, was taken.) Additional measures selected and recorded, but not included in this analysis, were yaw and pitch angle errors (in degrees) and rate errors (in degrees per second) in all axes.

Earth Entry (ES). The earth entry, as performed in this simulation, again imposed fairly complex requirements which were, in turn, reflected in a variety of measures. Initially, proper orientation had to be assumed and this was represented by attitude angle error (in degrees). Then for the first 210 seconds tracking of a roll program of ramps and flats was critical. Performance of this was again represented, like pitch tracking in the insertion phases, by the average and the standard deviation of error during the period. Next altitude and altitude rate errors and later crossrange and crossrange rate errors were controlled by appropriate deviations from the commanded roll program. Performance in these respects was considered best represented by average and standard deviation of error. Overall accuracy of the entry was represented by the terminal displacement from the landing site and efficiency represented by fuel consumed. All of these measures were considered and recorded as of special interest; however, again as a matter of economy and convenience, a still more restricted set was used in the following analysis. Thus, the measures used in this analysis are simply terminal displacement and altitude and crossrange errors (all in feet). The altitude and crossrange measures used were again the estimation (based on the respective mean and standard deviation of error) of the maximum error occurring with 95% frequency, as for pitch error in the orbital insertions. The criteria set were 50,000 feet displacement and 8,579.6 feet in both altitude and crossrange error (based on a mean of 2,000 and a standard deviation of 4,000 feet). Fuel consumption was not considered because, although quite possibly useful in differentiating among different levels of acceptable performance, it is of slight relevance to mission effectiveness in this phase.

Even though the number of measures was severely limited in the fashion described, there still remain 22 measures, distributed over 9 mission phases, to represent each pilot's flight control performance in the test mission. These are presumed to effectively represent performance aspects which are both critical to mission success and sensitive to differences in performance within and among individuals. These measures are summarized (with abbreviations) and the associated criteria are given in table V.

Analytic Methodology

That there were so many measures of pilot performance raised significant problems in analytic methodology. It is conventional in psychological research to limit the measures of performance (the dependent variable) to one or a few measures, taken repeatedly. This arrangement permits a determination of performance differences directly in terms of the measures used and any significant variation in performance and performance differences may readily be attributable to experimental manipulations in the variable under study. But, as applied to this circumstance, such a method would lead to, again, a total of 22 outcomes for each individual and comparison of interest—certainly an unmanageable number to interpret effectively in a direct manner. Furthermore, the 22 measures make reference to almost as many different ranges of variation in magnitude and 5 different units of measure and so cannot be directly combined without undesired bias. Nevertheless, it was imperative, as a study goal, that the data be interpreted in such a way as to permit statements concerning overall phase and, in turn, mission effectiveness—not just a collection of statements concerning performance in individual parameters. Also, it seemed very desirable to express these overall statements in terms of performance reliabilities (probabilities) rather than in terms of arbitrary units of measure. Expression in terms of reliability, or the language of design effectiveness, would enhance the meaningfulness of the results to a broader readership and make them more directly applicable to design problems.

Therefore, rather than to limit the treatment of the recorded data to a simple transformation into Z-scores (i.e., interpreting the raw data in terms of a normal density function having the same estimated parameter values), which is a common method for reducing multiple measures to a common basis for combination, a novel method was developed and employed. Essentially, this involved the interpretation of performance on each parameter at any stage of the study (training, test or retest) as an estimated probability of performing at a specified criterion value. These probabilities or reliabilities were then combined in each of two ways to arrive at somewhat different kinds of statements, all with unique and useful meaning, concerning performance reliability in the various phases and over the whole mission. Thus, as a result, it is possible to state, alternatively, the following about performance on any given occasion:

(1) the probability of meeting criterion performance in a measured parameter of a given mission phase; defined as

$$\bar{p} = \frac{\Sigma_i^n p}{n}$$

(2) the probability of meeting criterion performance in all measured parameters of a given mission phase, defined as

$$\dot{p} = \prod_{1}^{m} p$$

(3) the probability of meeting criterion performance in a measured parameter of the complete mission; defined as

$$\bar{\bar{p}} = \frac{\sum_{1}^{k} \sum_{1}^{m_k} p}{\sum_{1}^{k} m_k}$$

(4) the probability of meeting criterion performance in all measured parameters in a phase of the complete mission; defined as

$$\bar{\dot{p}} = \frac{\sum_{1}^{k} \prod_{1}^{m_k} p}{k}$$

(5) the probability of meeting criterion performance in all measured parameters of all phases of the complete mission; defined as

$$\dot{p} = \prod_{1}^{k} \prod_{1}^{m_k} p$$

In each of the defining expressions m is the number of measures in a phase and k is the number of phases in the mission.[1] Taken together these five indicants represent a hierarchy of inclusiveness. The first, $\bar{p}$, represents merely the average probability of success in a given parameter, as estimated on the basis of a given phase; whereas the last, $\dot{p}$, represents the probability of successfully performing the whole mission (i.e., not exceeding the criterion for any parameter).

Seen in another way, of course, the indicant $\dot{p}$ also represents the resultant of a particular model of pilot flight control within the simulated system. A relevant question which may be raised then is whether the simple model expressed by relationship 5 is an appropriate one for the present purpose. While not without problems, as noted later, this model did seem to be the best choice within the limits of present methodology. In fact, if the need is stringently expressed it may be considered the only choice possible. A general discussion of the methodological problem of modeling (or measuring) complex task performance and this approach to it is planned for early publication. Appendix II includes a description of the analytic details applicable to this study.

Finally, besides the problem of handling the different measures obtained in such a way as to arrive at useful statements concerning phase and mission reliabilities, there is yet another analytic problem posed by these data. It is simply that, for any performance comparison of interest, except certain comparisons of performance on the part of the same individual, there is only one unique measure available. Thus, with respect to retention interval each

---

[1] For readers unfamiliar with the other symbols, Σ stands for "the sum of" the item series indicated and Π stands for "the joint product of" the item series indicated.

(2) the probability of meeting criterion performance in all measured parameters of a given mission phase, defined as

$$\dot{p} = \prod_{1}^{m} p$$

(3) the probability of meeting criterion performance in a measured parameter of the complete mission; defined as

$$\bar{\bar{p}} = \frac{\sum_{1}^{k} \sum_{1}^{m_k} p}{\sum_{1}^{k} m_k}$$

(4) the probability of meeting criterion performance in all measured parameters in a phase of the complete mission; defined as

$$\bar{\dot{p}} = \frac{\sum_{1}^{k} \prod_{1}^{m_k} p}{k}$$

(5) the probability of meeting criterion performance in all measured parameters of all phases of the complete mission; defined as

$$\dot{p} = \prod_{1}^{k} \prod_{1}^{m_k} p$$

In each of the defining expressions m is the number of measures in a phase and k is the number of phases in the mission.[1] Taken together these five indicants represent a hierarchy of inclusiveness. The first, $\bar{p}$, represents merely the average probability of success in a given parameter, as estimated on the basis of a given phase; whereas the last, $\dot{p}$, represents the probability of successfully performing the whole mission (i.e., not exceeding the criterion for any parameter).

Seen in another way, of course, the indicant $\dot{p}$ also represents the resultant of a particular model of pilot flight control within the simulated system. A relevant question which may be raised then is whether the simple model expressed by relationship 5 is an appropriate one for the present purpose. While not without problems, as noted later, this model did seem to be the best choice within the limits of present methodology. In fact, if the need is stringently expressed it may be considered the only choice possible. A general discussion of the methodological problem of modeling (or measuring) complex task performance and this approach to it is planned for early publication. Appendix II includes a description of the analytic details applicable to this study.

Finally, besides the problem of handling the different measures obtained in such a way as to arrive at useful statements concerning phase and mission reliabilities, there is yet another analytic problem posed by these data. It is simply that, for any performance comparison of interest, except certain comparisons of performance on the part of the same individual, there is only one unique measure available. Thus, with respect to retention interval each

---

[1] For readers unfamiliar with the other symbols, Σ stands for "the sum of" the item series indicated and Π stands for "the joint product of" the item series indicated.

TABLE V

Measures Used in Analysis of Flight Control

Phase and Measure	Criterion
Translunar Insertion (TLI) *	
velocity cutoff error (V)	10 fps
pitch control error (P)	0.2645 deg
Transposition (TEN)	
displacement rate (DR)	1 fps
impact rate (XR)	1 fps
fuel consumed (F)	10 slugs
Lunar Orbit Insertion (LOI)	
velocity cutoff error (V)	10 fps
pitch control error (P)	0.2645 deg
Separation and Deorbit (SDO)	
velocity cutoff error (V)	2 fps
yaw control error (Y)	1 deg
pitch control error (P)	2 deg
Brake and Hover (BH)	
displacement (or range error) (D)	200 ft
displacement rate (DR)	10 fps
impact rate (XR)	10 fps
percentage fuel consumed (F)	95 %
Lunar Ascent (LA)	
velocity cutoff error (V)	10 fps
Rendezvous (Rend)	
percentage fuel consumed (F)	95 %
Docking (Dok)	
displacement (D)	1 ft
displacement rate (DR)	0.5 fps
impact rate (XR)	9.1 fps
Earth Entry (EE)	
displacement (or range error) (D)	50,000 ft
altitude error (A)	8,579.6 ft
crossrange error (C)	8,579.6 ft

* Also applicable to Transearth Insertions (TEI) performed by C-4 and C-9 in retraining.

of the crews was tested at a somewhat different interval and so may not (except possibly for C-8 and C-9) be defensibly combined. Rather evident differences in the adequacy of original training in the two phases of the study (as will be shown) argue against combining C-8 and C-9, even if 8- and 9-week intervals are otherwise considered the same for the purpose of this study.

Furthermore, within each group the crewmembers were tested in a unique order on the supposition (borne out by the data) that those tested second and third might benefit from participating in the preceding tests, even though not acting as pilot. In addition, as is also suggested by the data on performance at the end of training, each pilot came to the test having previously acquired a unique level of skill on the several mission tasks not necessarily paralleling the level of skill achieved by any other one. Yet, it is known that previous level of skill may strongly influence amount of measured skill retention. Therefore, the performances of the several pilots within a given crew may not be combined for cross-comparison purposes and used for obtaining error estimates basic to testing the significance of differences without incurring the risk of selective bias resulting from the interactions operating.

The net consequence of these facts is that the study data must be viewed and the results consistently interpreted with reference to the rather complex data structure illustrated in table VI. In the presentation of results to

TABLE VI

Structure of the Data

Crew/Retention Interval		Stages of Experiment		
		Training	Test	Retraining
C-4 - Pilot/Order	1			
	2			
	3			
C-9 - Pilot/Order	1			
	2			
	3			
C-8 - Pilot/Order	1			
	2			
	3			
C-13 - Pilot/Order	1			
	2			
	3			

follow the values obtained for the several crewmembers of a given crew have been combined, on occasion. But, the dubiety of this procedure, in view of the previous, should be constantly borne in mind. Conventional tests of significance were not considered appropriate and the interpretation of the results must rest upon the observed internal and logical consistency of the data. The situation is frankly describable, from a statistical point of view, as a zero degrees of freedom case. Nevertheless, the nature of the data permits the drawing of some useful conclusions.

SECTION III

RESULTS

In keeping with the previous statement on methodological considerations, the main results of the study pertinent to flight control are presented in this section in terms of probabilities (or reliabilities)—first with reference to skill at the conclusion of training, next with reference to retention test performance, and finally with reference to retraining performance. In all cases the probabilities reported are rounded to the nearest .001 with the result that a probability of .0005 or less is reported as <.001 and treated as zero in any calculations and a probability of .9995 or greater is reported as >.999, or simply indicated by a dash in tables, and treated as one in any calculations. By way of limiting the amount of tabular material to that most immediately useful, the basic values presented are the probabilities per phase of success in any parameter ($\bar{p}$) or in all parameters of the phase ($\hat{p}$). These may be thought of, respectively, as the average probability and the joint probability per phase. More inclusive indicants for individual pilots and crews are given also. However, the elemental probabilities for each measured parameter (of the 22 described earlier), from which $\bar{p}$, $\hat{p}$ and the still more inclusive indicants are derived, may be found in the tables of appendix III for reference.

Skill Level Attained in Training

The first question which must appropriately be asked of the study data is naturally, What level of skill was achieved by the test personnel prior to the retention period and subsequent test? Research on skill retention has frequently demonstrated that level of prior learning strongly influences measured skill retention. Therefore, it is important to assess the relative level of the several individuals and crews involved in this test in order to validly interpret their test performance. Otherwise, for example, differences or lack of differences easily attributable to differences in learning might be falsely ascribed to differences in duration of retention. In this case assessment of prior levels seems especially important because the impressions of study personnel as well as the relative numbers of recorded training experiences in the various mission phases clearly suggest that the training received by C-8 and C-13 was superior. Furthermore, the desire to generalize the study findings to planning for space operations implies that the prior skill achievement of the test personnel should be demonstrably high. For, it may be assumed that crews of space systems will continue (at least for some time to come) to be trained to the near limits of their skill potential.

Reliability Per Hypothetical Criteria. One approach to assessing prior skill achievement is to simply compare the best performance of the pilots, presumably occurring at the end of their training, with the hypothetical system criterion for each of the several measures of interest. If the criteria selected do represent reasonable or typical mission requirements, as is supposed, then such a comparison will show the capability of the pilots to perform the mission. Moreover, since the performance of each individual and crew would be assessed relative to a common reference, any differences in achievement should become evident.

The results of such a comparison are summarized in table VII. In this table are listed for each pilot the phase-by-phase reliabilities at the end of training, as estimated from the last four training trials (with the Fast Time and Real Time Mission excluded). (See appendix II for details concerning computation of these values.) In each case the two primary values, $\bar{p}$ and $\dot{p}$, are given. These primary values are then taken cumulatively across phases to obtain the values $\bar{\bar{p}}$ and $\dot{\bar{p}}$ and jointly to obtain $\dot{p}$ for the complete mission for each pilot. Finally, the $\bar{p}$ and $\dot{p}$ values for phases and the $\bar{\bar{p}}$, $\dot{\bar{p}}$ and $\dot{p}$ values for the mission for the three individual pilots within a crew are taken cumulatively to obtain the values of $\bar{\bar{p}}$ and $\dot{\bar{p}}$ for phases and $\bar{\bar{\bar{p}}}$, $\dot{\bar{\bar{p}}}$ and $\dot{\bar{p}}$ for the mission for each crew. Logically, these last five reliabilities indicating crew performance merely state the most likely reliability to be expected, of the nature otherwise indicated, if a member of that particular crew performs the particular phase or the mission (when performance by the members is equally likely). This method of tabulation will be repeated in other tables as well.

Even rather cursory examination of table VII suggests that there were differences in skill at the end of training which are worthy of note. First, for all phases the value of $\dot{p}$ does not exceed the value of $\bar{p}$ and generally (except where only one measure is involved) it is smaller. This is to be expected from the difference in algebraic manipulation defining each. Hence, evidently and by definition $\dot{p}$ is the more sensitive indicant, such that the lowest value of $\dot{p}$ obtained for any pilot in any phase is .002 (P-41 in EE) whereas the lowest value of $\bar{p}$ obtained is .158 (C-91 in LA). Both indicants range upward in a number of instances to >.999.

Second, evidently there was considerable variation in the pilots' capability to perform the various mission phases. Thus, for example, $\bar{p}$ for P-41 varied from a high of >.999 in the LOI to a low of .259 for EE, $\dot{p}$ ranging from >.999 to .002 for the same phases. Similarly P-91 ranged from .816 to .158 in $\bar{p}$ and .632 to .158 in $\dot{p}$. However, in contrast P-81 and P-131 showed much less individual variation. There is a suggestion of similarity among pilots in relative effectiveness on the several phases, especially among those of C-4 and C-9 where EE is generally performed less well, but there seems to be no such entirely consistent pattern. Individual differences as well as selective differences in training might easily have contributed to this result.

Third, it is evident also that so far as overall mission capability at the end of training is concerned the pilots differed considerably. Thus, $\bar{\bar{p}}$ ranged from .587 to .991 ($\dot{\bar{p}}$ from .312 to .978 and $\dot{p}$ from <.001 to .747) within this sample of 12. Once again, of course, both differences in training as well as in skill potential may have contributed to this result.

Finally, the mission capabilities of the pilots comprising the several crews clearly suggests that pilots of C-4 and C-9 were much less skilled at the end of training than were pilots of C-8 and C-13. This difference is clearly confirmed by the reliability estimates for the phases and the mission for each whole crew. Thus, the mission reliabilities, $\dot{\bar{\bar{p}}}$, for C-4 and C-9 were .726 and .648 as contrasted with .984 and .953 obtained for C-8 and C-13 and comparable differences exist for $\bar{\bar{p}}$ and $\dot{\bar{p}}$ values as well. An illustration of these differences between C-4 and C-9 when combined and C-8 and C-13 when combined is given in figure 7 which depicts overall $\bar{\bar{p}}$ and $\dot{\bar{p}}$ values. When the information on the rendezvous is eliminated in computing the values for C-8 and C-13 to provide a common 8-phase basis for comparison the differences are even greater.

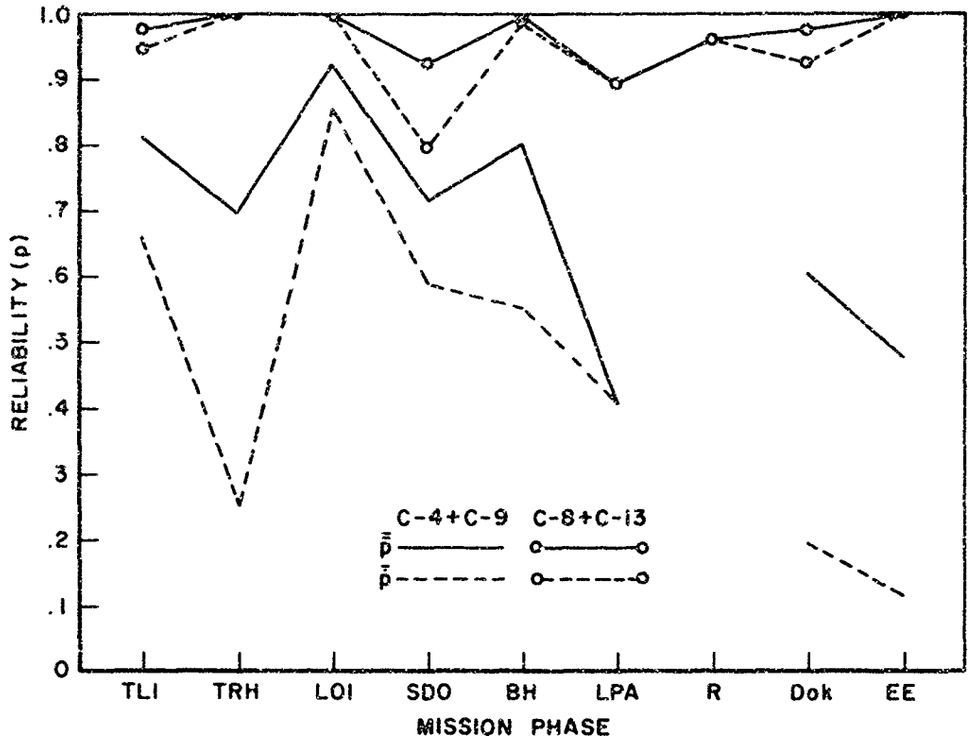


Figure 7. Phase Reliabilities of Combined Crews at the End of Training

TABLE VII

Reliability Per Criteria at the End o. Training

Crew/ Pilot		TL1	THN	LO1	SDO	Phase BH	LA	Rend	Dok	EE	Mission $\bar{p},\bar{p}$	Mission $\dot{p}$
C-4	1 $(\bar{p})$	.892	.745	—†	.928	.450	.358	ND	.726	.259	.670	
	1 $(\dot{p})$	.785	.234	—	.785	.030	.358		.228	.002	.428	<.001
	2	.957	.547	.798	.827	.999	.345	ND	.289	.535	.662	
	2	.914	.054	.600	.480	.997	.345		.022	.129	.443	<.001
	3	.928	.668	—	.864	.995	—	ND	.751	.554	.845	
	3	.860	.016	—	.593	.981	—		.392	.107	.619	<.001
	all $(\bar{p})$▽	.926	.653	.933	.873	.815	.568	ND	.589	.449	.726	
	all $(\dot{p})$	.853	.101	.867	.619	.669	.568		.214	.079	.497	<.001
C-9	1 $(\bar{p})$	.816	.614	.761	.472*	.618	.158	ND	.637	.619	.587	
	1 $(\dot{p})$	.632	.170	.545		.159	.158		.186	.176	.312	<.001
	2	.534	.917	.996	.339*	.813	ND	ND	.723	.654	.711	
	2	.276	.752	.992		.395			.220	.256	.461	.002
	3	.750	.691	—	.803*	.937	.170	ND	.506	.243	.646	
	3	.500	.287	—		.755	.170		.116	.003	.462	<.001
	all $(\bar{p})$	.700	.741	.919	.560*	.789	.164	ND	.622	.505	.648	
	all $(\dot{p})$	.469	.403	.846		.436	.164		.174	.145	.412	.001
C-8	1 $(\bar{p})$	.990	—	—	.919	—	—	—	—	—	.990	
	1 $(\dot{p})$	.979	—	—	.757	—	—	—	—	—	.971	.741
	2	—	—	—	.953	—	.806	—	.996	—	.973	
	2	—	.999	—	.860	.999	.806	—	.988	—	.961	.683
	3	.954	—	.992	.999	—	—	—	.972	—	.991	
	3	.908	—	.983	.998	—	—	—	.917	—	.978	.817
	all $(\bar{p})$	.981	—	.997	.957	—	.935	—	.989	—	.984	
	all $(\dot{p})$	.962	—	.994	.872	—	—	—	.968	—	.970	.747
C-13	1 $(\bar{p})$	—	—	.999	.932	.980	.999	.789	—	—	.967	
	1 $(\dot{p})$	—	—	.999	.804	.920	.999	.789	—	—	.946	.582
	2	—	—	—	.739	—	.999	.978	.939	.998	.961	
	2	—	—	—	.370	—	.999	.978	.819	.994	.907	.294
	3	.903	—	—	.995	.998	.557	.997	.942	—	.932	
	3	.806	.999	—	.984	.993	.557	.997	.827	—	.907	.361
	all $(\bar{p})$	.968	—	—	.889	.993	.852	.921	.960	.999	.953	
	all $(\dot{p})$	.935	—	—	.719	.971	.852	.921	.882	.998	.920	.412

† dash indicates value >.9995

* based on reduced data

ND data we.. unavailable

▽ probabilities for "all" pilots are individual values taken cumulatively (i.e., means of individual values)

Those apparent and large differences between the crews of the two test phases and the nature of the reliabilities obtained, when viewed in the context of the preceding observations concerning level of learning, cast serious doubt upon the value of the data from the first two crews tested for supporting any useful generalizations on retention. Certainly, any generalizations concerning the performance of C-4 and C-9 must be duly related to the evidently less than complete training received by these crews, for there is no evident basis for contending that they were inherently less capable.

Capability Relative to Skill Potential. However, that C-8 and C-13 were by far the more capable crews and reached rather high levels of reliability according to arbitrary criteria does not establish that even they had reached anywhere near the limits of their skill potential. Instead, to determine their actual level of skill relative to their potential what is needed is some way of determining when improvement in measured skill has stopped and/or a method of specifying levels of learning (i.e., training) short of or beyond that level. It has often been demonstrated that practice beyond the point of further measured improvement still brings continued learning (technically, called over-learning) as evidenced by improved retention. But, the traditional means for attempted satisfaction of this need are neither standardized nor adequate. Thus, commonly the assessment of asymptotic performance rests upon a personal judgment of the investigator as to whether the mean performance taken over successive blocks of trials is reasonably stationary. Levels of learning (or training) are usually specified in terms of actual practice time or practice trials, relative to practice or trials required for asymptotic performance level, as estimated from the same data (when the asymptote was reached) or like data (when the asymptote was not reached). To specify asymptotic performance by means of a personal judgment is, of course, to invite inaccuracy in the specification. To specify level of training short of the asymptote on a personalized crude extrapolation from typical data and the terminal rate observed is to court still greater inaccuracies. Furthermore, to specify the asymptote or the level in terms of the mean performance is to discount the variation in performance from instance to instance in the face of the commonly held belief that good performance is not only typically good, but reliable as well. On the other hand, if the performance changes indicated by even the means of successive performance samples are considerable and consistently in the direction of improvement it may certainly be concluded that skill has not yet reached peak levels.

The data on training and retraining performance in the present study were thus examined with an eye to roughly estimating how near the test personnel approached their skill limits at the end of training. However, in making this analysis and in keeping with the study goals primary attention was given to the variation in estimated performance level for successive performance samples which, from the data, were estimated as achievable with a high degree of reliability. Thus, instead of considering the successive levels of arithmetic mean (or presumably most typical performance) the successive levels of performance estimated (from four-trial blocks) to include 95% of performances under exactly the same circumstances were examined. In this way changes in variability as well as in typical performance level were represented. Although 95% admittedly does not represent an extremely high reliability it does represent an effective compromise between degree of certainty and increasing errors in estimating the associated performance level. The logic for this approach to performance measurement is described in a separate article by Cotterman [ ].

Viewed in this way the training and retraining data from C-4 and C-9 simply confirm the previously mentioned conclusion about their end-of-training skills. There is great variation in estimates of $p_{.95}$ levels for successive four-trial blocks in training and in most instances there is still a large trend toward improvement. The comparable or superior $p_{.95}$ levels displayed in retraining, which will be presented in more detail later, provide further confirmation.

However, the estimated $p_{.95}$ levels in training for C-8 and C-13 do not permit such a clear, unambiguous conclusion. For any given individual the series of estimates seems to have generally stabilized for some parameters and not to have stabilized for others, even within a phase, and this pattern is not consistent among individuals. The stability noted is generally to be found at the very end of training and to involve only a limited number (2 or 3) four-trial blocks. Examples of these training functions (and retraining, as well) are given in appendix IV in which the data from P-131 are depicted graphically. In these graphs the estimated $p_{.95}$ level is coded by $\dot{X}$ to distinguish it from the mean, or $\bar{X}$, and the recorded value on a given trial, or X, in traditional coding. A reasonable interpretation of these results would be simply that by the end of training the several pilots involved neared in varying degrees on the several parameters, but had not quite reached (with the exception of certain parameters) the limits of their skill potential. This interpretation also is supported by the retraining results to be presented in detail later.

Further confirmation of this evaluation of the end-of-training skill of C-8 and C-13 may be derived (depending on personal judgment) from the limited application of a newly-developed decision rule to their training data. This training decision rule provided a means for more precisely determining when asymptotic skill level was achieved for the purpose of distributing the available training time for greatest overall training effectiveness. However, when applied to the data in question in only a few cases did the outcome result in a finding of skill stabilization (i.e., peak skill attainment). It seemed apparent that available training time would not permit skill stabilization, at least as indicated by the decision rule used. Since additional calculations not considered directly relevant to analysis with reference to system criteria were involved, with attendant costs, the study manager abolished the requirement for its application. The decision rule concept is described in a general way elsewhere (Cotterman, 1967) but it has not been elaborated and properly tested. Therefore, being as yet unproven the rule may be questioned as to validity and, in the form used, as to excessive stringency. For the moment, then, the interpretation of the outcomes obtained with it seems necessarily a matter of personal judgment.

Nevertheless, whatever the validity of the decision rule, because stability in estimated $p_{.95}$ performance level would certainly constitute a severe requirement by usual standards it seems fair to view C-8 and C-13 as having neared, but not quite reached in training the limits of their potential skill. Such, although never precisely determined, is not unlikely the result of most operational training programs. Accordingly, the data of C-8 and C-13 are taken provisionally as a reasonable basis for generalizations concerning skill retention and, in the following description of mission test and retraining performance, they are considered of primary significance. The parallel analyses for C-4 and C-9 are presented, but they are considered of lesser value.

Retention Test Mission Performance

Reliability per Hypothetical Criteria. How well then did the crews perform in the test mission of main concern? Since end-of-training performance has been expressed as sets of estimated reliabilities with reference to 22 flight control criteria, an evident approach to answering this question is to compare with them similar estimates based on performance in the test mission. The performance effects of the retention periods should then be reflected in differences between end-of-training and test mission reliability estimates. Hence, the $\bar{p}$ and $\dot{p}$ values per phase for each pilot were computed, as before, and the more inclusive $\bar{\bar{p}}$, $\bar{p}$ and $\dot{p}$ values for each individual and the parallel cumulative probabilities for crews determined.[2] All these values are presented in table VIII, which is organized in the same fashion and may be compared directly with table VII concerning end-of-training performance.

First, taking the values of table VIII strictly on their own merits, it is evident that the pilots showed considerable individual variation in capability to perform the several mission phases in the test mission as at the end of training. Among the commanders (those first tested), for example, $\bar{p}$ for P-81 varied from >.999 in a number of phases to .684 in TL1, and $\dot{p}$ from >.999 in the same several phases to .138 in SDO. The variation in phase performance of P-131 is even more extreme, ranging from >.999 to .294 in $\bar{p}$ and >.999 to <.001 in $\dot{p}$. Such variation in phase performance is also typical of all the pilots of C-4 and C-9 but it did not occur in the performance of the C-8 and C-13 pilots who were tested second and third. The more consistent and superior performance of the second and third pilots of C-8 and C-13 may be the result of their greater skills (compared with those of C-4 and C-9) and their opportunity to observe the first tested pilots before performing the test mission themselves. Again, there was a tendency for all pilots, especially within a given crew, to perform certain phases better than others, but the pilots were not entirely consistent in this respect.

In fact, when the phase reliabilities for the several pilots within a given crew are compared there is considerable agreement on some and considerable variation on others. Almost invariably in C-8 and C-13, the performance of the first tested pilot was least adequate, suggesting rather definitely the effects of order of test among those about equally and well trained individuals. That similar relations are not as evident among the test performances of C-4 and C-9 pilots is not surprising in view of the incompleteness of their training and greater variation in their terminal skill.

---

[2] Because no direct estimate of variability could be made from the single measure, in calculating test mission probabilities the variability at the end of training was used. (See appendix II for details.) In general, it may be expected that this procedure resulted in an underestimate or overestimate of test mission variability, depending on whether performance in the test mission was better or worse than end-of-training performance. Thus, considering the nature of these data, it is supposed that the test mission reliabilities reported are usually overestimates. This matter is more fully discussed in the following section.

TABLE VIII

Reliability Per Criteria in the Test Mission

Crew/ Pilot		Phase TLl	TEN	LOl	SDO	EH	LA	Rend	Dok	EE	Mission ▽ $\bar{p}$,$\dot{p}$	$\dot{p}$
C-4	1 ($\bar{p}$)	—†	.752	—	.531	.350	.635	ND	.187	.737	.649	
	($\dot{p}$)	—	.255	—	<.001	.099	.635	ND	<.001	.305	.412	<.001
	2	.999	.828	.919	.530	.998	.655	ND	.654	.418	.750	
		.998	.526	.838	<.001	.994	.655	ND	.277	.024	.539	<.001
	3	.992	.705	—	.826	.861	—	ND	.762	.380	.816	
		.983	.116	—	.497	.506	—	ND	.412	.005	.565	<.001
	all ($\bar{p}$)▽	.997	.762	.973	.629	.736	.763	ND	.534	.512	.738	
	($\dot{p}$)	.994	.299	.946	.166	.533	.763	ND	.230	.111	.505	<.001
C-9	1 ($\bar{p}$)	.964	.243	.946	ND	.250	.583	ND	.640	.740	.624	
	($\dot{p}$)	.928	<.001	.893	ND	<.001	.583	ND	.188	.382	.425	<.001
	2	.752	.692	.958	.517*	.250	ND	ND	.540	.793	.643	
		.556	.075	.916	.517*	<.001	ND	ND	.092	.462	.374	<.001
	3	.898	.830	—	.843*	.635*	.557	ND	.443	.747*	.744	
		.797	.536	—	.843*	.130*	.557	ND	.080	.521*	.558	.001
	all ($\bar{p}$)	.871	.588	.968	.680	.378	.570	ND	.541	.760	.670	
	($\dot{p}$)	.760	.204	.936	.680	.043	.570	ND	.120	.455	.452	<.001
C-8	1 ($\bar{p}$)	.684	—	—	.713	—	—	—	.999	.981	.931	
	($\dot{p}$)	.367	—	—	.138	—	—	—	.998	.942	.827	.048
	2	.750	—	—	.980	.968	.966	—	—	—	.963	
		.500	—	—	.939	.871	.966	—	—	—	.920	.395
	3	.993	—	.992	.990	—	—	—	.982	.998	.995	
		.986	—	.983	.970	—	—	—	.946	.995	.987	.885
	all ($\bar{p}$)	.809	—	.997	.894	.989	.989	—	.994	.993	.963	
	($\dot{p}$)	.618	—	.994	.682	.957	.989	—	.981	.979	.911	.443
C-13	1 ($\bar{p}$)	—	—	—	.960	.294	.903	.936	—	—	.899	
	($\dot{p}$)	—	—	—	.880	<.001	.903	.936	—	—	.858	<.001
	2	.996	—	—	.893	—	.964	—	.992	.997	.982	
		.993	—	—	.695	—	.964	—	.975	.992	.958	.643
	3	.961	.999	—	.976	.981	.716	—	.951	—	.954	
		.922	.996	—	.928	.925	.716	—	.853	—	.927	.481
	all ($\bar{p}$)	.986	—	—	.943	.758	.861	.979	.981	.999	.945	
	($\dot{p}$)	.972	.999	—	.834	.642	.861	.979	.943	.997	.914	.375

† dash indicates value >.9995

* based on reduced data

ND data were unavailable

▽ probabilities for "all" pilots are individual values taken cumulatively (i.e., means of individual values)

The overall mission performance of the pilots as indicated by $\bar{\bar{p}}$ clearly and consistently demonstrates these effects of testing order; within each crew the least reliable was the first tested. The more stringent indicants, $\bar{p}$ and $\dot{p}$, also show this effect to the extent that a probability of sufficient magnitude (that is, .0005) for discriminating differences was found. Furthermore, in general the performance of the C-8 and C-13 pilots was considerably superior to that of the C-4 and C-9 pilots, as would be expected from the differences in skill at the end of training which are reflected in table VII.

However, when the mission reliabilities in table VIII for individuals and for crews are contrasted with the comparable values of table VII some loss in reliability over the retention interval seems evident. For example, the estimated reliabilities for P-131 decreased from .967 to .899 in $\bar{\bar{p}}$, from .946 to .858 in $\bar{p}$ and from .582 to <.001 in $\dot{p}$ from training to the test mission. Only one of the four pilots first tested (P-91) did not show an overall loss and that this pilot did not may easily be considered the result of inadequate skill in the first place ($\bar{\bar{p}}$ of .587 at the end of training). Graphs illustrating the losses in reliability indicated by test performance are given in figure 8 in which the phase reliabilities ($\dot{p}$) of P-81 and P-131 in training and in the test mission are plotted. As crews, both C-8 and C-13 show overall losses even though certain individuals showed slight gains. That C-4 and C-9 did not show overall losses may again be attributed to the inadequacy of their training. Because the pilots of these crews had gained less skill originally, they had relatively less skill to lose.

These relationships between overall mission performance in training and test are more conveniently and precisely summarized in table IX which shows both the absolute amount of change in reliability from training to test and the percentage of end-of-training reliability that change represents. Thus, according to table IX the first tested pilots, whose performance is of primary interest because it is uncontaminated by prior participation, never lost more than .068 or 7% of end-of-training reliability in $\bar{\bar{p}}$. Because $\bar{p}$ and $\dot{p}$ are more stringent indicants, the decrements in them are greater, as much as .144 or 14.8% and .582 or 100%, respectively. What these figures mean, of course, is that while cumulative (or average) parameter reliabilities held up rather well in the test mission the likelihood of mission success (as defined by meeting all hypothetical criteria) dropped greatly. This degradation in $\dot{p}$ is invariably the result of degradation in only one or a few parameters, as may be seen more directly by comparing tables XXV and XXVI of appendix III.

The shifts from training to test in parameter ($\bar{\bar{p}}$) and phase ($\bar{p}$) reliabilities for the whole mission are not only relatively small but they also do not seem to be ordered consistently in magnitude according to retention interval. It would be expected that if other factors are equalized the longer the retention interval the greater the loss (or lesser the gain) in reliability. Among the first tested pilots (the test of whom was uncontaminated), while P-131 did show the expected slightly greater loss than P-81 and P-81 greater loss than P-41, the gain of P-91 is completely out of order. Thus, instead of P-41<P-81<P-91 <P-131 the order is P-91<P-41<P-81<P-131. That P-91 lost nothing (in fact, gained) in test may again be explained on the grounds of his comparative lack of skill at the end of training; that is, he had comparatively little to lose. However, even granting this explanation, because the differences observed are slight it can hardly be concluded that the length of the retention interval had much effect. There appears little basis in these results for arguing that a

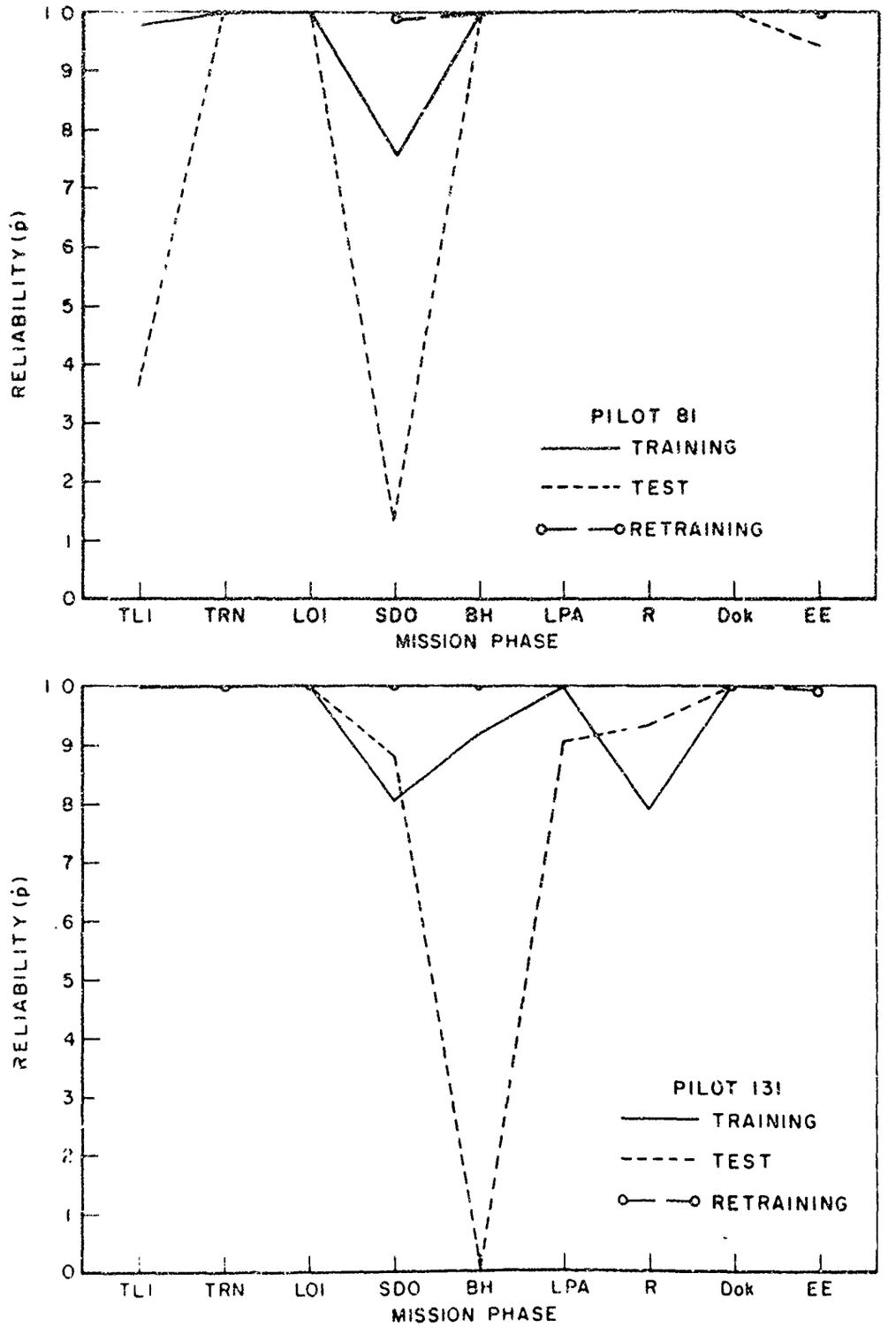


Figure 8. Phase Reliabilities of P-81 and P-131 in Training, in the Test Mission, and in their Best Four-Trial Block of Retraining

TABLE IX

Change in Mission Reliability per Criteria from Training to Test Mission *

Crew/Measure		Pilot 1	2	3	All (mean)
4	$\bar{\bar{p}}$	-.021 (-3.1)	.088 (13.3)	-.029 (-3.4)	.013 (2.3)
	$\bar{p}$	-.016 (-3.7)	.096 (21.7)	-.054 (-8.7)	.009 (3.1)
	$\dot{p}$	NM	NM	NM	NM
9	$\bar{\bar{p}}$	.037 (6.3)	-.068 (-9.6)	.098 (15.2)	.022 (4.0)
	$\bar{p}$	.113 (36.2)	-.087 (-18.9)	.096 (20.8)	.041 (12.7)
	$\dot{p}$	NM	-.002 (-100.0)	NM	NM
8	$\bar{\bar{p}}$	-.059 (-6.0)	-.010 (-1.0)	.004 (0.4)	-.022 (-2.2)
	$\bar{p}$	-.144 (-14.8)	-.041 (-4.3)	.009 (0.9)	-.059 (-6.1)
	$\dot{p}$	-.693 (-93.5)	-.288 (-42.2)	.068 (8.3)	-.304 (-42.5)
13	$\bar{\bar{p}}$	-.068 (-7.0)	.021 (2.2)	.022 (2.4)	-.008 (-0.8)
	$\bar{p}$	-.088 (-9.3)	.051 (5.6)	.020 (2.2)	-.006 (-0.5)
	$\dot{p}$	-.582 (-100.0)	.349 (118.7)	.120 (33.2)	-.038 (17.3)

* percentage gain or loss indicated in parenthesis
NM not meaningful because training reliability <.0005

13-week retention period has much greater significance for subsequent performance than a 4- or 8-week retention period. The average shifts of the crews, being ordered C-9<C-4<C-13<C-8, support this same view.

However, as the overall reliability shifts of table VII for pilots within given crews are compared it is once again evident that excepting P-91 those tested first tended to perform less well. Certainly this is true of C-8 and C-13 for which no obvious question of training adequacy can be raised. That the second and third tested individuals performed so well, relatively, suggests a considerable training (i.e., retraining) value accrued to them merely from observing and/or aiding one or two prior performances.

Capability Relative to Previous Level. Thus, to the extent that the hypothetical criteria are representative of those existing in real space missions and the analytic methodology employed affords valid estimates of reliability a number of useful inferences concerning human reliability in space operations seem possible from the preceding analysis. These inferences will be developed more explicitly and discussed, as warranted, in the following section of the report. Meanwhile, in focusing attention in the analysis so far presented on the several pilots' and crews' capabilities relative to criteria, some information (as it happens in this case) has been lost on the full effects of the retention period. This circumstance results from the fact that at the end of training the pilots' reliabilities with respect to many of the measured parameters far exceeded the maximum discriminable value (.9995) adopted as practicable in this analysis. Therefore, an individual might well have performed less adequately in test without this fact being reflected in the calculated probability. That would always be true when capability exceeds the maximum discriminable reliability relative to a particular measure. In effect, then, to compute reliabilities relative to criteria in these circumstances is to reduce the capability of the analysis to show certain changes in performance. Nevertheless, in this study and in many others, one is properly interested in any changes in performance which occur—not just in those changes which have obvious and immediate implications for operational performance.

Therefore, as a means of assessing more fully the nature of the performance in the test mission, contrasted with that of training, an additional analysis of the test data was performed. In this analysis, (again in keeping with the view that performance excellency is indicated by the level that can be achieved reliably, not just typically), reference is made to the estimated level of performance which the individual would achieve in 95% of his performances under like circumstances (i.e., his $p_{.95}$ level). Thus, with respect to each flight control parameter of interest, the $p_{.95}$ level of performance estimated from the last four training trials was used as the reference against which the test mission performance was compared. As in the previous analysis the actual performance in test was used as a basis for estimating the likelihood of achieving this criterion (the $p_{.95}$ level in this case). Hence, equivalent performance in test is indicated by a probability of .950 (one equal to the reference), superior performance by a probability greater than .950 and inferior performance by a probability less than .950. Also, as before, the probabilities for separate parameters were taken cumulatively to obtain $\bar{p}$ and taken jointly to obtain $\dot{p}$ values for each of the several phases. These, in turn, also were combined as before to obtain still more inclusive values for individuals performing the whole mission and for crews. However, since the estimated probability of achieving the criterion in this case is .950 the

expected value for a given $\dot{p}$ is something less than .950, as a direct function of the number of separate probabilities which are thereby taken jointly. For example, the expected $\dot{p}$ for a phase represented by three measures is .857. $\bar{p}$ remains unaffected and the expected value for it is always .950.

The primary results of this analysis of test mission data may be seen in table X which, like the parallel table VIII, summarizes the $\bar{p}$ and $\dot{p}$ values obtained by phase and the more inclusive indicants based upon these. (The probabilities for each individual and each parameter are tabled in appendix III.) Table X shows that the likelihood of pilots in C-8 and C-13 achieving their end-of-training $p_{.95}$ level in the test mission was nowhere near as great as their likelihood of achieving the hypothetical criteria. This result is to be expected in view of their great likelihood of meeting the criteria at the end of training. But, in contrast, the pilots of C-4 and C-9 had about the same likelihood of reaching their $p_{.95}$ levels as they did of reaching the criteria—again to be expected in view of their lesser skill at the end of training.

Once again, as was noted in the prior analysis, the individual pilots showed considerable variability in their capability to perform in test at former levels on the several phases. This is illustrated by the two examples in figure 9—again depicting the performances of P-81 and P-131, but now with reference to the $p_{.95}$ level established by each individually rather than with reference to a common arbitrary criterion. Thus, both these pilots showed considerable reduction in performance of certain phases and not of others. However, they were not entirely consistent in this. Although both had difficulty with braking and hover, docking, and earth entry, P-81 also had difficulty with the translunar insertion and the separation and deorbit, whereas P-131 had difficulty with transposition. This pattern of variability in phase-to-phase performance and of only moderate consistency among pilots in phase performance is typical. Even the pilots of C-8 and C-13 who were tested second and third show phase-by-phase variations in test performance relative to former levels, when they did not relative to criteria. This may be considered confirmatory evidence of the greater sensitivity to performance changes that analysis by reference to former individual capability brings.

The overall mission performance probabilities shown in table X also clearly indicate a loss on the part of all individuals and crews in the test mission, over capability at the end of training. Furthermore, the amount of loss seems again to be related to the order of test—the first tested individual always having performed less adequately relative to his previous skill than the others. These relationships can be noted more conveniently from table XI in which is listed for each pilot the amount of loss in probability from the expected value based on $p_{.95}$ training levels and the percentage of the expected value such a loss represents. (All values in this table indicate losses and so are not signed.) Thus, for example, among the first tested pilots deterioration in test from individual $p_{.95}$ level expectancies ranged from .228 to .165 (or 24 to 17.4%) in $\bar{p}$. Their losses in $\dot{p}$, of course, are even greater—from 46.3 to 22.3%. All showed a probability of <.0005 of achieving their expected $\dot{p}$ value and, hence, a nominal 100% loss in it. However, as already noted, the losses of the second and third tested pilots are never so great (except with respect to $\dot{p}$ which, being so low in any case, does not discriminate among them). The natural result is that the mean losses of the crews are somewhat less than those of the first tested pilots—ranging from 17.8 to 11.6% in $\bar{p}$ and 39.0 to 23.9% in $\dot{p}$. Regarding these overall losses (both of the first tested pilots

TABLE X

Probability Estimated from Test Mission Performance of Attaining the $p_{.95}$ Skill Level Achieved in Training

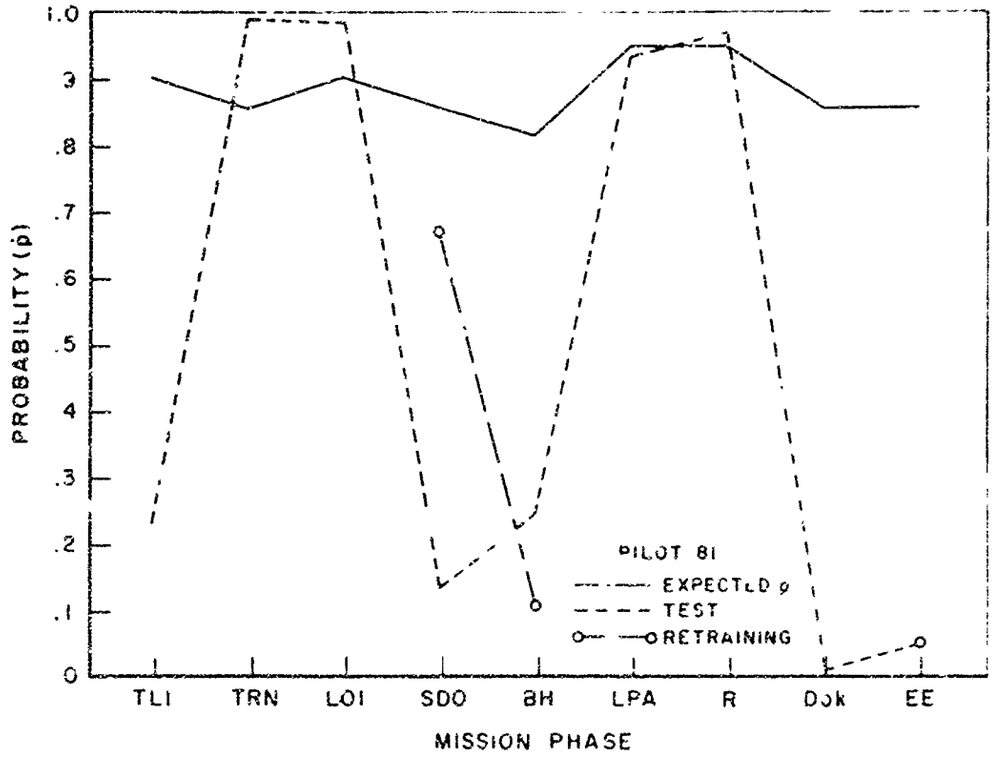
Crew/ Pilot		Phase TLl	TRN	LOI	SDO	BH	LA	Rend	Dok	EE	Mission $\bar{p}, \dot{p}$	Mission $\dot{p}$
C-4	1 ($\bar{p}$)	—†	.587	.988	.591	.672	.991	ND	.331	.994	.769	
	($\dot{p}$)	—	<.001	.977	<.001	.001	.991	ND	<.001	.981	.494	<.001
	2	.999	.967	.992	.640	.463	.993	ND	.994	.862	.864	
		.998	.903	.985	<.001	<.001	.993	ND	.981	.630	.686	<.001
	3	.995	.975	—	.803	.621	—	ND	.966	.731	.886	
		.990	.926	.999	.504	.113	—	ND	.900	.384	.727	.018
	all ($\bar{p}$)▽	.998	.843	.993	.678	.585	.995	ND	.764	.862	.840	
	($\dot{p}$)	.996	.610	.987	.168	.038	.995	ND	.627	.665	.636	.006
C-9	1 ($\bar{p}$)	.998	.330	.997	ND	.250	.998	ND	.967	.956	.785*	
	($\dot{p}$)	.997	<.001	.994	ND	<.001	.998	ND	.903	.870	.680*	<.001*
	2	.988	.769	.862	.982*	.250	ND	ND	.822	.977	.807*	
		.977	.314	.725	.982*	<.001	ND	ND	.494	.932	.632*	<.001*
	3	.497	.989	.998	.931*	.494*	.997	ND	.963	— *	.859*	
		<.001	.671	.996	.931*	<.001*	.997	ND	.892	— *	.686*	<.001*
	all ($\bar{p}$)	.828	.696	.952	.956*	.331	.998*	ND	.917	.978*	.817*	
	($\dot{p}$)	.658	.328	.905	.956*	<.001	.998*	ND	.763	.934*	.666*	<.001*
C-8	1 ($\bar{p}$)	.614	.997	.992	.534	.723	.933	.969	.394	.442	.733	
	($\dot{p}$)	.231	.991	.984	.134	.244	.933	.969	.010	.048	.505	<.001
	2	.452	.963	.827	.933	.431	.995	—	.979	.659	.804	
		.001	.891	.657	.802	.023	.995	—	.938	<.001	.590	<.001
	3	.994	.965	.888	.553	.698	.998	.997	.653	.520	.807	
		.989	.896	.786	<.001	.005	.998	.997	<.001	.007	.520	<.001
	all ($\bar{p}$)	.687	.975	.902	.673	.617	.975	.989	.675	.540	.781	
	($\dot{p}$)	.407	.926	.809	.312	.091	.975	.989	.316	.018	.538	<.001
C-13	1 ($\bar{p}$)	—	.591	.984	.853	.285	.382	.991	.653	.755	.722	
	($\dot{p}$)	—	<.001	.968	.600	<.001	.382	.991	.026	.301	.474	<.001
	2	.712	.939	.966	.979	.946	.957	—	.996	.644	.871	
		.424	.819	.931	.937	.797	.657	—	.989	.025	.731	.004
	3	.634	.954	.768	.764	.558	.981	.983	.885	.710	.805	
		.278	.893	.572	.362	.046	.981	.983	.689	.322	.570	.001
	all ($\bar{p}$)	.782	.831	.906	.865	.596	.673	.991	.845	.703	.799	
	($\dot{p}$)	.567	.571	.824	.633	.281	.673	.991	.568	.216	.592	.002

† dash indicates value >.9995

* based on reduced data

ND data were unavailable

▽ probabilities for "all" pilots are individual values taken cumulatively (i.e., means of individual values)



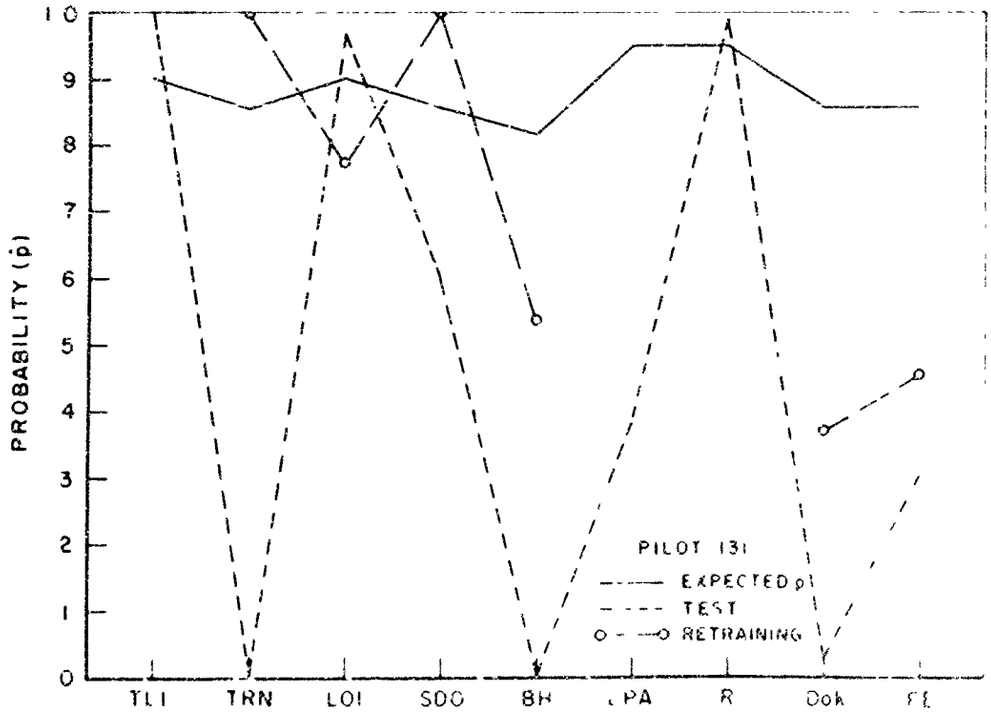


Figure 9. Expected and Obtained Phase Probabilities of P-81 and P 131 Achieving Their p.95 Levels of Training in the Test Mission and in their Last Four-Trial Block of Retraining

TABLE XI

Loss in Test Mission Reliability
from $p_{.95}$ Training Level *

Crew/Measure		Pilot 1	Pilot 2	Pilot 3	All (mean)	p (expected)
4	$\bar{\bar{p}}$	.181 (19.1)	.086 (9.1)	.064 (6.7)	.110 (11.6)	.950
	$\bar{p}$	.381 (43.5)	.189 (21.6)	.148 (16.9)	.239 (27.3)	.875
	$\dot{p}$	.340 (100.0)	.340 (100.0)	.322 (94.7)	.334 (99.2)	.340
9	$\bar{\bar{p}}$	.165 (17.4)	.143 (15.1)	.091 (9.6)	.133 (14.0)	.950
	$\bar{p}$	.195 (22.3)	.243 (27.8)	.189 (21.6)	.209 (23.9)	.875
	$\dot{p}$	.340 (100.0)	.340 (100.0)	.340 (100.0)	.340 (100.0)	.340
8	$\bar{\bar{p}}$	.217 (22.8)	.146 (15.4)	.143 (15.1)	.169 (17.8)	.950
	$\bar{p}$	.378 (42.8)	.293 (33.2)	.363 (41.1)	.345 (39.0)	.885
	$\dot{p}$	.323 (100.0)	.323 (100.0)	.323 (100.0)	.323 (100.0)	.323
13	$\bar{\bar{p}}$	.228 (24.0)	.079 (8.3)	.145 (15.3)	.151 (15.9)	.950
	$\bar{p}$	.409 (46.3)	.152 (17.2)	.313 (35.4)	.291 (33.0)	.883
	$\dot{p}$	.323 (100.0)	.319 (98.8)	.322 (99.7)	.321 (99.5)	.323

* percentage loss indicated in parentheses

and the average losses of the crews) it is of special interest that the differences among crews and individuals is quite small. This is in contrast with the previously described results with reference to criteria in which both C-4 and C-9 showed average gains and in which one first tested individual actually showed a gain. Thus, since apparently considerable difference in skill level existed, it may be said that whatever the level of skill acquired the deterioration in reliability from that level or relative to it was nearly constant.

On the other hand, that little difference among individual and/or crew losses was found argues again that the effects of retention interval, if any, were slight. The first tested pilots' losses in reliabilities are ordered $91<41<81<131$ in both $\bar{p}$ and $\tilde{p}$ while those of the crews are ordered $4<9<13<8$ in $\bar{p}$ and $9<4<13<8$ in $\tilde{p}$. Apparently, although the differences are small and somewhat inconsistent, loss in reliability did tend to be directly related to retention interval.

In general, the results of this additional analysis of test performance which makes reference to individual capability, parallel those of the first, which makes reference to hypothetical system criteria. The principal differences in outcome which can be noted are (1) that the losses in capability relative to end-of-training $p_{.95}$ levels are about two to three times greater and (2) that the findings for C-4 and C-9 are not in complete agreement, as would be expected on the basis of their less adequate training. Otherwise, both analyses clearly indicate a degree of degradation in test mission performance compared with end-of-training performance, varying considerably by phase for individual pilots in a not completely consistent fashion, but which overall phases show relatively evident effects of the order in which the pilots were tested and relatively slight effects of the retention interval involved.

Retraining Performance

Having shown that the pilots did experience a loss in skill over the retention period the next question is, How well did they regain their skill in the subsequent retraining trials? The rapidity with which they achieved their former levels, if they did achieve them, would carry valuable implications as to the possibilities of mitigating the degrading effects of lengthy retention periods in operational circumstances which warrant the concern. Furthermore, because even the pilots of C-8 and C-13 had not necessarily reached the limits of their skill potential in training (those of C-4 and C-9 definitely had not) it is of special interest to ascertain if, possibly, in the course of the retraining given them they would not only regain, but surpass, their former levels. In fact, it was just this possibility that prompted the 3 days of retraining for C-8 and C-13.

With these two questions in view, the analysis of retraining performance paralleled that of test mission performance, focusing attention first on estimated likelihood of reaching the hypothetical system criteria and next on estimated likelihood of achieving the level of performance achievable 95% of the time at the end of original training. Because (with the exception noted) pilots of C-8 and C-13 had as many as 28 and never fewer than 8 retraining trials on certain phases it was possible to consider their data in more than one way. Thus, the analysis just mentioned was first performed upon their

initial four-trial block of retraining data and then repeated upon the best four-trial block for each individual and phase ($\hat{p}$ was the criterion used in selecting the best trial block). In contrast, because the pilots of C-4 and C-9 received much less retraining (from 8 to 2 trials, depending on phase) and the data for these trials are not quite complete all their available retraining data were considered in a single analysis. This analysis is considered roughly comparable to the analysis of initial four-trial performance of C-8 and C-13 pilots.

Reliability per Hypothetical Criteria. Accordingly, the results of the analysis of initial retraining performance of all pilots relative to criteria are summarized together in table XII. As in previous tables of this nature the $\bar{p}$ and $\hat{p}$ values for separate phases are indicated, followed by the more inclusive probabilities referring to all phases and the whole crew. The phase reliabilities are, of course, directly comparable to those reported in table VII for original training and in table VIII for the test mission. However, the all-phase probabilities for individuals and crews are not directly comparable to the mission values of the former tables, because they are based on a more limited set of phases (namely, those which had been selected for retraining).

Judging from the phase probabilities reported in table XII, as before, the pilots showed varying capabilities to perform the several phases. They also varied considerably among each other in capability to perform any particular phase—especially the pilots of C-4 and C-9. However, together they showed a rather consistent tendency to perform certain phases better than others. This is clearly evident in the values for C-8 and C-13 pilots—none of whom apparently had difficulty with the transposition, lunar orbit insertion, and docking phases. But as a group they apparently had some difficulty with the separation and deorbit, braking and hover, and earth entry phases. This pattern is reflected also in the lower crew probabilities for C-4 and C-9 for the same phases.

In general, the phase probabilities, both for individuals and for crews, are quite high, approximating those of training (table VII) and surpassing those of the test mission (table VIII), as would be expected. Once again the performance levels of the C-8 and C-13 pilots was distinctively superior to that of the C-4 and C-9 pilots. Furthermore, when the best four-trial performance of C-8 and C-13 pilots, shown in table XIII, is considered it is seen that additional opportunities for training resulted in even greater performance reliability. In more than two-thirds of the phases (23 out of 33) these pilots met or exceeded their end-of-training levels. A graphic illustration of retraining performance may be seen by referring to figure 8 which includes a plot of $\hat{p}$ values obtained for P-81 and P-131, along with the comparable plots representing training and test mission performance.

Going beyond the pilots' phase-by-phase performance, a more complete view of their retraining achievements may be had by examining tables XIV (for C-8 and C-13) and XV (for C-4 and C-9). In the left-hand columns in each of these tables are summarized the $\bar{\bar{p}}$, $\bar{p}$ and $\hat{p}$ values with respect to criteria for individuals, and the related cumulative probabilities for crews, in training, in the test mission, and in retraining. The probabilities given for training and test mission in these tables were calculated on the basis of only those phases on which retraining was given and are, therefore, directly comparable to the probabilities for retraining. (Having been computed on the basis of only certain mission phases, they are not the same values as are given in tables VII and VIII.)

TABLE XII

Reliability Per Criteria in Early Retraining **

Crew/ Pilot		Phase: TRN (2)	LOI/ TRI(2)	SDO	BH (8)	Dck (6)	BE (6)	All Phases ▽ $\bar{p}, \dot{p}$	$\dot{p}$
C-4 1	$\bar{p}$	.818	.990	ND	.443	.894	.388	.707	
	$\dot{p}$	.455	.979	ND	.033	.683	<.001	.430	<.001
2		.668	.702	ND	.723	.598	.645	.667	
		.144	.404	ND	.241	.213	.054	.211	<.001
3		.938	.996	ND	.907	.978	.985	.959	
		.817	.991	ND	.627	.934	.956	.865	.453
all	$\bar{p}$ ▽	.808	.896	ND	.691	.823	.673	.778	
	$\dot{p}$	.472	.791	ND	.300	.610	.337	.502	.151
C-9 1	$\bar{p}$	.724	.766	ND	.485	.731	.550	.651	
	$\dot{p}$	.326	.587	ND	.048	.353	.118	.286	<.001
2		.817	.665	ND	.616	.881	.863	.768	
		.452	.414	ND	.100	.680	.631	.455	.008
3		.765	.930*	ND	.554	.661	.538	.689*	
		.402	.930*	ND	.079	.285	.141	.367*	.001*
all	$\bar{p}$	.769	.787*	ND	.552	.758	.650	.703	
	$\dot{p}$	.393	.644*	ND	.076	.439	.297	.369	.003
C-8 1	$\bar{p}$	ND	ND	.740	—†	ND	.692	.811*	
	$\dot{p}$	ND	ND	.231	—	ND	.284	.505*	.066*
2		—	—	.990	.998	—	—	.998	
		—	—	.969	.991	—	—	.993	.960
3		—	—	—	—	—	.976	.996	
		—	—	—	—	—	.928	.988	.928
all	$\bar{p}$	—*	—*	.910	.999	—*	.889	.935*	
	$\dot{p}$	—*	—*	.733	.997	—*	.737	.829*	.651*
C-13 1	$\bar{p}$	—	—	—	.996	—	.992	.998	
	$\dot{p}$	—	—	—	.982	—	.977	.993	.959
2		—	—	.979	—	—	—	.996	
		—	—	.937	—	—	—	.990	.937
3		—	—	.847	.953	—	.736	.923	
		—	—	.546	.811	—	.336	.782	.149
all	$\bar{p}$	—	—	.942	.983	—	.909	.972	
	$\dot{p}$	—	—	.828	.931	—	.771	.922	.682

** for C-8 and C-13 based on the first four training trials; for C-4 and C-9 based on all retraining trials, as noted in parentheses

▽ probabilities for "all" pilots are individual values taken cumulatively

† dash indicates value >.9995

* based on reduced data

TABLE XIII

Reliability of C-8 and C-13 per Criteria as Estimated from Past Four-Trial Block in Retraining

Crew/ Pilot		Phase TRN	LOI	[illegible]	NH	Dok	EE	All Phases ▼ $\bar{p}, \dot{p}$	$\dot{p}$
C-8	1 ($\bar{p}$)	ND	ND	.995	—†	ND	.997	.997*	
	($\dot{p}$)	ND	ND	.986	—	ND	.992	.993*	.978*
	2	—	—	—	—	—	—	—	
		—	—	—	—	—	—	—	—
	3	—	—	—	—	—	.976	.996	
		—	—	—	—	—	.928	.988	.928
	all ($\bar{p}$)▼	—*	—*	.998	—	—*	.991	.998*	
	($\dot{p}$)	—*	—*	.995	—	—*	.973	.994*	.969*
C-13	1 ($\bar{p}$)	—	—	—	—	—	.995	.999	
	($\dot{p}$)	—	—	—	—	—	.985	.998	.985
	2	—	—	.979	—	—	—	.996	
		—	—	.937	—	—	—	.990	.937
	3	—	—	.991	.953	—	—	.991	
		—	—	.973	.811	—	—	.964	.789
	all ($\bar{p}$)	—	—	.990	.984	—	.998	.995	
	($\dot{p}$)	—	—	.970	.937	—	.995	.984	.904

† dash indicates value >.9995

* based on reduced data

▼ probabilities for "all" pilots are individual values taken cumulatively (i.e., means of individual values)

TABLE XIV

Comparative Reliabilities of Crews C-8 and C-13 in Training, Test and Retraining **

Crew/Pilot		Per Criteria in: Trng (last 5)	Mission Test	Retrng (1-4)	Retrng (best 4)	Per $p_{.95}$ Trng Level in: Mission Test	Retrng (1-4)	Retrng (best 4)
C-8 1	$\bar{p}$	.986	.949	.811*	.997*	.566	.509*	.644*
	$\tilde{p}$	.960	.846	.705*	.993*	.142	.036*	.275*
	$\dot{p}$	.757	.130	.066*	.978*	.002	.001*	.004*
2		.992	.991	.998	.999	.799	.828	.863
		.974	.968	.993	.999	.552	.515	.624
		.848	.818	.960	.999	.001	.002	.025
3		.994	.994	.996	.996	.713	.800	.829
		.983	.982	.988	.988	.282	.466	.631
		.900	.898	.928	.928	.001	.001	.005
all (mean)	$\bar{p}$	.991	.978	.935* (.966)†	.998*	.693 (.730)	.712* (.799)	.779* (.827)
	$\tilde{p}$	.972	.932	.829* (.911)	.994*	.385 (.412)	.339* (.461)	.510* (.603)
	$\dot{p}$	.835	.615	.651*	.969*	.001	.001*	.011*
C-13 1	$\bar{p}$	.985	.876	.998	.999	.687	.850	.892
	$\tilde{p}$	.954	.813	.993	.998	.316	.585	.686
	$\dot{p}$	.739	.001	.959	.985	.001	.002	.068
2		.946	.980	.996	.996	.912	.884	.969
		.864	.944	.990	.990	.750	.690	.902
		.301	.672	.937	.937	.014	.053	.521
3		.989	.984	.923	.991	.775	.815	.951
		.967	.950	.782	.964	.481	.630	.837
		.807	.729	.149	.789	.002	.004	.288
all (mean)	$\bar{p}$	.973	.947	.972	.995	.791	.850	.937
	$\tilde{p}$	.928	.902	.922	.984	.516	.635	.808
	$\dot{p}$	.616	.467	.682	.904	.005	.020	.292

** based on the six phases of retraining (i.e., TRN, LOI, SDO, EH, DOK, EB)
Note: expected values for $\bar{p}$, $\tilde{p}$, and $\dot{p}$ per $p_{.95}$ level are .950, .858, and .397

* based on incomplete data

† parenthetical values obtained by cumulating crew values for phases, instead of pilot values for mission. They differ because data for P-81 are incomplete

TABLE XV

Comparative Reliabilities of Crews C-4 and C-9 in Training, Test and Retraining **

Crew/Pilot			Per Criteria in: Trng (last 4)	Mission Test	Retrng	Per $p_{.95}$ Trng Level in: Mission Test	Retrng
C-4	1	$\bar{p}$	.614	.605	.707	.714	.878
		$\bar{p}$	.256	.332	.430	.392	.646
		$\hat{p}$	<.001	<.001	<.001	<.001	.084
	2		.674	.763	.667	.856	.782
			.440	.532	.211	.700	.546
			<.001	.003	<.001	<.001	.001
	3		.788	.742	.959	.859	.888
			.489	.408	.865	.664	.658
			.001	<.001	.453	.036	.025
	all (mean)	$\bar{p}$	.692	.703	.778	.810	.849
		$\bar{p}$	.395	.424	.502	.585	.617
		$\hat{p}$	<.001	.001	.151	.012	.037
C-9	1	$\bar{p}$	.611	.564	.651	.700	.876
		$\bar{p}$	.198	.293	.286	.553	.680
		$\hat{p}$	<.001	<.001	<.001	<.001	.046
	2		.698	.647	.768	.736	.859
			.339	.309	.455	.493	.649
			.001	<.001	.008	<.001	.035
	3		.665	.731*	.689*	.889	.861*
			.422	.453*	.367*	.712	.731*
			<.001	.003*	.001*	<.001	.027*
	all (mean)	$\bar{p}$	.658	.647	.703	.775	.865
		$\bar{p}$	.320	.352	.369	.586	.687
		$\hat{p}$	<.001	.001	.003	<.001	.036

** based on the five phases of retraining (i.e., TRN, HH, DOK, TRI, MM) with LOI substituted for TRI in the test mission. Expected values for $\bar{p}$, $\bar{p}$, and $\hat{p}$ per $p_{.95}$ level are .950, .858, and .463, respectively

* based on incomplete data

From table XIV it is evident that so far as the hypothetical system criteria are concerned, the pilots of C-8 and C-13 quickly regained their overall effectiveness during the retraining provided. Four of the six had already exceeded their end-of-training levels in their first four retraining trials and all virtually met or exceeded them in their best four-trial block of retraining. The cumulative probabilities for both crews naturally reflect this also by indicating gains in the best four trials with respect to all aspects ($\bar{\bar{p}}$, $\bar{p}$, $\dot{p}$) of overall effectiveness. Some individuals—for example, P-82, P-131 and P-132—showed especially noteworthy improvement. Also, it is likely that had not P-81 been somewhat disturbed by the requirement to terminate his participation after just one-half day of retraining he would have performed still more effectively, with the result of greater demonstrated gain for himself and his crew. Of course, the estimated reliabilities of all the pilots were very high in the best four-trial block in retraining—$\bar{\bar{p}}$ ranging from .999 to .991, $\bar{p}$ from .999 to .964, and $\dot{p}$ from .999 to .789. That these reliabilities reflect some gain over original training implies that, as was surmised, these pilots had not quite reached the limit of their skill potential at the end of training.

The overall effectiveness in retraining of C-4 and C-9, as shown in table XV, runs parallel to that of C-8 and C-13 in some respects and in others it does not. The pilots of C-4 and C-9 (P-42 excepted) also regained their former levels of reliability in the brief retraining afforded them. Some, notably P-41, P-43, and P-92, even showed substantial gains. Similarly, the cumulative probabilities for the two crews reflect this general improvement in all aspects of overall effectiveness. However, the estimated reliabilities with respect to criteria of C-4 and C-9 pilots at their best were still generally low when compared with those of the C-8 and C-13 pilots. $\bar{\bar{p}}$ ranged from .959 to .651, $\bar{p}$ from .865 to .286 and $\dot{p}$ from .453 to <.001. Clearly, most of the C-4 and C-9 pilots had not even approached the limits of their skill potential in original training and did not do so in retraining either.

Capability Relative to Previous Level. However, when the retraining performance of the pilots/crews is examined in terms of the likelihood of achieving their end-of-training $p_{.95}$ level a somewhat different, although generally complementary, view is obtained. The phase-by-phase probabilities of achieving this former level are summarized for the initial retraining of all pilots in table XVI and for the best four-trial blocks of C-8 and C-13 pilots in table XVII. (Once again, although the $\bar{p}$ values to be expected are .950 the $\dot{p}$ values to be expected vary with the number of measured parameters for the phase, as in the comparison table X concerning test mission performance.) In addition, the directly comparable figures for overall effectiveness in the test mission and the expected values may be seen in the right-hand columns of tables XIV and XV, for C-8 and C-13 and C-4 and C-9, respectively.

The phase probabilities for initial retraining given in table XVI are often, but certainly not always, higher than those for the test mission given in table X. In fact, only on the braking and hover and docking phases are the cumulative probabilities for crews consistently higher. Furthermore, the probabilities based on initial retraining seldom reach the expected value (.950, etc.) thus indicating that the individuals and crews performed few phases initially at their former levels of skill. None of them attained their former level in separation and deorbit, braking and hover, and docking.

Again, there was considerable variation within individuals in their capability to perform the several phases at their former level and among individuals in their capability to perform a given phase at former levels. Overall, in contrast with the analysis in terms of criteria, their effectiveness on the several phases does not show a very consistent pattern of effectiveness favoring certain phases over others, although pilots of C-8 are still seen to have performed less competently on the same three phases identified earlier. In general, the phase-by-phase probabilities of achieving the former $p_{.95}$ level do not seem to vary as greatly as did those of meeting the hypothetical criteria.

When contrasted with the entries of table XVI those of table XVII show, as in the criterial analysis, the effects of the additional retraining received by the pilots of C-8 and C-13. As crews in the best four trial block of retraining they surpassed their initial four-trial levels in nine out of twelve instances. However, even these most skilled pilots at their best did not consistently exceed their training level and in several phases were far short of it. This is illustrated in figure 9 in which retraining performances of P-81 and P-131 are plotted with reference to their former $p_{.95}$ levels. The result, in overall effectiveness, was that only one of the six (P-132) slightly exceeded the expected value and one other approximately matched that value (see table XIII). The discrepancy between the values for P-81 and those of the other pilots again may be most easily attributed to the adverse circumstances under which he performed. From these findings it may be argued that the pilots of C-8 and C-13 had indeed reached their skill potential at the end of training—a view which seemingly contradicts the earlier findings on performance with respect to criteria.

Having noted, then, that C-8 and C-13 did not much surpass their former levels it naturally might be supposed that the less-skilled C-4 and C-9 were even further from their former levels in retraining. But, the last column of table XIV shows that this was not so, that if anything, C-4 and C-9 surpassed C-8 and C-13 in capability to perform at former levels in the initial retraining. This observation too is at variance with the findings on retraining performance with respect to criteria.

In general, concerning retraining it is concluded that in the first few trials most pilots regained considerable skill—enough to demonstrably surpass their test mission capability—but not enough to match their end-of-training levels. The amount by which they fell short of the $p_{.95}$ training level (from .250 to .061 in $\bar{p}$, excluding P-81) was surprisingly similar, irrespective of their degree of original skill. Those receiving additional retraining approached and, in at least a third of the cases, matched their former levels but showed little tendency to surpass them. End-of-training capability to meet hypothetical system criteria was fully regained by virtually all the pilots in a few retraining trials, several pilots actually surpassing it, and was always surpassed in the best retraining performance of those receiving additional retraining. But, the less well-trained pilots of C-4 and C-9 continued in retraining to be distinctively less able to meet the criteria than the pilots of C-8 and C-13, even though they performed equally well relative to their own former levels.

Evidently then, the first question as to how quickly skill was reacquired may best be answered in this way: both skilled and relatively unskilled pilots were able to regain their capability to meet the hypothetical system criteria within about four trials following the test mission, but not their full original capability. The originally well-trained pilots did approximate or equal their

TABLE XVI

Probability Estimated from Early Retraining Performance of Attaining the $p_{.95}$ Skill Level of the Last Four Training Trials **

Crew/ Pilot		Phase TRE (2)	LOI/ TEI(2)	SDO	BE (8)	Dok (6)	EE (6)	All Phases ▽ $\bar{p}, \dot{p}$	$\dot{p}$
C-4	1 ($\bar{p}$)	.820	.956	ND	.825	.825	.958	.878	
	($\dot{p}$)	.534	.911	ND	.363	.543	.878	.646	.084
	2	.905	.506	ND	.520	.999	.979	.782	
		.718	.022	ND	.053	.998	.938	.546	.001
	3	.688	.970	ND	.824	—†	.960	.888	
		.077	.939	ND	.396	—	.879	.658	.025
	all ($\bar{p}$)▽	.804	.811	ND	.723	.941	.966	.849	
	($\dot{p}$)	.443	.624	ND	.271	.847	.898	.617	.037
C-9	1 ($\bar{p}$)	.983	.978	ND	.628	.924	.867	.876	
	($\dot{p}$)	.950	.955	ND	.108	.774	.611	.680	.046
	2	.905	.992	ND	.602	.835	.959	.859	
		.733	.984	ND	.102	.549	.878	.649	.035
	3	.940	.933*	ND	.482	—	.950	.861	
		.827	.933*	ND	.041	—	.855	.731	.027
	all ($\bar{p}$)	.943	.968	ND	.571	.920	.925	.865	
	($\dot{p}$)	.837	.957	ND	.087	.774	.781	.687	.036
C-8	1 ($\bar{p}$)	ND	ND	.581	.564	ND	.383	.509*	
	($\dot{p}$)	ND	ND	.029	.027	ND	.051	.036*	<.001*
	2	.913	.832	.689	.734	.822	.978	.828	
		.755	.663	.069	.134	.534	.933	.515	.002
	3	.997	—	.667	.655	.878	.602	.800	
		.990	—	.022	.033	.665	.087	.466	<.001
	all ($\bar{p}$)	.955*	.916*	.646	.651	.850*	.654	.712*	
	($\dot{p}$)	.872*	.832*	.040	.065	.600*	.357	.339*	.001*
C-13	1 ($\bar{p}$)	.999	.886	—	.694	.788	.930	.850	
	($\dot{p}$)	.997	.771	—	.021	.366	.356	.585	.002
	2	.962	.744	.949	.923	—	.728	.884	
		.887	.502	.846	.704	—	.199	.690	.053
	3	.999	.934	.789	.874	—	.296	.815	
		.996	.869	.391	.500	—	.025	.630	.004
	all ($\bar{p}$)	.987	.855	.913	.830	.929	.585	.850	
	($\dot{p}$)	.960	.714	.746	.408	.789	.193	.635	.020

** for C-8 and C-13 based on the first four training trials; for C-4 and C-9 based on all retraining data, as noted in parentheses

▽ probabilities for "all" pilots are individual values taken cumulatively

† dash indicates value >.9995

* based on reduced data

TABLE XVII

Probability Estimated from Best Four-Trial Block in Retraining of C-8 and C-13 Attaining the $p_{.95}$ Skill Level of the Last Four Training Trials

Crew/ Pilot		Phase TR'	LOI	SDO	BH	Dok	EE	All Phases ▽ $\bar{p},\bar{p}$	$\dot{p}$
C-8	1 ($\bar{p}$)	ND	ND	.885	.664	ND	.383	.644*	
	($\dot{p}$)	ND	ND	.668	.107	ND	.051	.275*	.004*
	2	.999	.832	.800	.715	.854	.978	.863	
		.997	.663	.401	.173	.574	.933	.624	.025
	3	—†	—	.482	.950	.920	.623	.829	
		—	—	.048	.804	.771	.161	.631	.005
	all ($\bar{p}$)▽	—*	.916*	.722	.776	.887*	.661	.779	
	($\dot{p}$)	.998*	.832*	.372	.361	.672*	.382	.510*	.011*
C-13	1 ($\bar{p}$)	.999	.886	—	.883	.788	.796	.892	
	($\dot{p}$)	.997	.771	—	.534	.366	.451	.686	.068
	2	.962	.986	—	.936	—	.932	.969	
		.887	.972	—	.757	—	.799	.902	.521
	3	.999	.999	.940	.874	—	.893	.951	
		.996	.999	.821	.500	—	.705	.837	.288
	all ($\bar{p}$)	.987	.957	.980	.898	.929	.874	.937	
	($\dot{p}$)	.960	.914	.940	.597	.789	.652	.808	.292

† dash indicates value >.9995
* based on reduced data
▽ probabilities for "all" pilots are individual values taken cumulatively (i.e., means of individual values)

full original capability with varying amounts of additional retraining in the several phases. In contrast, the second question as to whether the pilots of C-8 and C-13 were fully trained appears unanswerable on the basis of the bare results so far presented. For, on the one hand, if capability to meet system criteria is the deciding factor they must be considered not to have been fully trained originally. On the other hand, if capability to perform at a highly reliable level ($p_{.95}$) determined from their own performance is the deciding factor there is no basis for supposing that they were not fully trained.

In the following section of the report this problematical aspect of the findings along with certain other aspects of the study which bear on the interpretation of the data will be discussed and some definitive conclusions will be stated.

SECTION IV

INTERPRETATION

The problem of interpreting the data obtained and the analytic results presented actually involves consideration of at least three somewhat different kinds of factors. These are, first of all, the obvious procedural factors relating to the manner of obtaining the data. Lack of standardization in procedure or failure to control or randomize some significant variable affecting the performance of the test personnel can certainly bias results. In order to interpret the study, then, it is desirable to consider the possibilities for bias in this regard and to gauge their likely impact on the data.

Beyond the details of data collection another aspect of the study significantly affecting the interpretation of findings concerns the details of the analytic procedures used. It is possible that some characteristic(s) of the analytic method employed may operate to introduce a bias in the representation or comparison of performance observed—thus resulting in an artifact, or spurious effect. Any possibilities of this nature should also be considered and their implications for the results determined.

Finally, beyond these procedural details of data collection and analysis there are always the somewhat broader questions of generalizability—of extent to which the findings are applicable to other situations. Although generalizability—depends greatly on the specific procedures used, it also depends upon more general matters relating to the very conceptualization of the study. These, too, should be examined, in retrospect, in order to arrive at a fully considered interpretation of the results.

Each of these three kinds of factors will be discussed, in turn, in the following portions of this section. Finally, in keeping with these interpretative remarks and the analytic results, specific study findings will be stated. Readers who do not care to follow the detailed examination of the numerous interpretative factors involved may turn directly to page 91 for the statement of findings. Alternatively, just the more specific considerations of data collection and analysis procedures may be skipped by turning to page 84 on which the discussion of generalizability begins.

Data Collection Procedures

Missing Data. With a study of the magnitude of this one in sheer volume of data it is a common occurrence that some of the data are lost—usually for a variety of reasons. Hence, it is a fair question to ask regarding this one, To what extent were data lost and how do these losses affect the results? This question can be answered in a very general way but of more use is specific information on where the losses were encountered.

Starting then with the training data, and restricting concern as always to flight control, the records of training performance exhibited by Grodsky et al (1966b) seem quite complete for the purpose of this analysis. Thus, as also may be noted in table XXV of appendix III (present report), sufficient data

on C-8 and C-13 were available to always provide a good four-trial basis for estimating end-of-training performance. In fact, with these two crews in only two cases (P-83 on TLI and P-131 on LOI) was there an evident difficulty requiring attention and this involved possibly erroneous, not missing data. In these two instances the last four values included one which was obviously extreme, as compared with the whole training record of the individual. Consequently, to obtain a fair estimate of capability these two values were replaced by reaching one trial further back in the training record. There were a few other instances where values which seemed unrepresentative of the individual's end-of-training capability were noted but, since they were not obviously extreme relative to the whole record they were used. Perhaps the most notable such instances are those of P-132 on SDO (velocity and yaw) and on Dok (displacement rate) and P-133 on LA (velocity). In these instances it is quite possible that the estimate of training is an underestimate and the comparisons of training with other performance are consequently biased. These possibilities will be explored more precisely in the discussion of sampling. Meanwhile, anticipating that discussion, it is doubtful that beyond affecting the estimates of original capability of P-132 and P-133 the extreme values entered greatly in the analysis of C-8 and C-13 performance.

In contrast, the record for C-4 and C-9 is not nearly so complete although, considering their evident lack of training and the secondary interest to be accorded their data, it still seems quite adequate. It was necessary sometimes to reach back more than four trials in training to secure a basis for estimation or to accept less than four trials as a basis, but this did not occur to a great extent. As table XXV indicates, acceptable estimates were possible for all but the r ndezvous performance, the SDO performance in yaw and pitch of C-9, and the LPA performance of P-92. Rendezvous performance cannot be estimated merely because fuel nsumption for it was not at that time being distinguished from that for docking in the performance records. This lack of information on rendezvous does force discounting rendezvous in the record of C-8 a l C-13 before comparison of the two sets of crews (as noted previously), but it otherwise has little bearing on the results. The other losses mentioned also seem of minor concern--merely a nuisance in calculation and in cross-comparison.

The test mission data are, in a sense, even more complete than the training data. Only the rendezvous information for C-4 and C-9 (for the reason already noted), the velocity control of P-91 in SDO and the displacement of P-93 in F[illegible] are unavailable. However, the lack of information on training performance o. C-9 in yaw and pitch in SDO and of P-92 in LPA as a basis for estimating variability, additionally made calculation of test mission probabilities for these performances impossible as well. Again though, considering the secondary interest in C-4 and C-9, the net effect of these data losses seems slight.

Similar general conclusions concerning the retraining data appear appropriate. For C-8 and C-13 the only losses were in the performance of P-81. These losses were, of course, considerable and they restrict opportunities to make fruitful comparisons involving the retraining performance of this especially significant first-tested pilot. But the sheer loss in data is probably not the main concern (as will be brought out more fully in a following discussion), and the principal interest in retraining is with respect to given individual's gains or losses over other performances. There are, of course, no data for retraining performance of C-4 and C-9 on SDO and LOI (TEI being substituted) which makes for nuisance in cross-comparison, but these are difficulties arising by design

rather than by less, as such. There are instances (appendix table XXVIII) in which the estimates for pilots of C-4 and C-9 had to be based on reduced data owing to a lack of information on a few parameters and trials and one case (P-93 on TEI, V) for which no estimate was possible.

Thus, in general, although the instances of missing data noted did cause some nuisance in analysis and do make cross-comparison more complicated than with a complete record, they certainly do not make meaningful interpretation impossible or, if properly handled, bias the results. Perhaps the most serious of these considerations is the problem (to be discussed in more detail) of possible bias due to the inclusion of unrepresentative values—but this problem is certainly not unique to this study.

Conduct of the Test. Apart from the broader issues of simulation involved, there are a number of particulars of the way in which this test was conducted which might possibly have influenced the results. In the first place, in order to make efficient use of personnel time and minimize costs involved, the full 7-day mission was compressed into a single work day by eliminating long coast periods and navigational tasks. In fact, it was compressed even more than that by arranging for three performances of each main phase—each of the three crew-members occupying each crew position once. What was the probable effect of such severe compression on the performances of the pilots relative to what they might have done? This question is naturally difficult to answer and perhaps frankly ought to be resolved on the basis of comparison data. However, lacking such data and yet being appreciative of the possible significance of the compression factor some estimation appears desirable. What then, in terms of known psychological variables, would be the likely effects of time compression?

In view of the available information on work periods and the effects of continuing work it readily may be supposed that as the test day wore on the pilots became somewhat less able to perform. This downward trend would be expected to continue until, toward the very last portion, the anticipation of completing the predefined task resulted in an end spurt. Similar trends, but of shorter duration, might also be expected within the morning until lunch was provided to the pilots in the simulated crew compartment and, possibly, within a given mission cycle for C-8 and C-13, the pilots of whom completed a whole mission in the same position. As the net result of such effects all the pilots (of all crews), as individuals, might be expected to exhibit parallel trends downward within the mission sequence until the last phase, with a slight improvement in any phase performed just before lunch. (For C-4 and C-9 the phase performed before lunch was LOI but for C-8 and C-13 the first-tested pilot completed EE just before lunch.) However, overlaid upon these predictable main trends are a number of virtually indeterminate possible effects on individuals as might arise from happenstances in scheduling, which afforded break periods in sporadic fashion and having to wait out relatively long periods of inactivity within the mission sequences of C-8 and C-13. (In fact, one C-13 pilot was observed showing all the typical signs of great boredom while waiting his turn in midafternoon.)

On the other hand, it is also known that such trends as these are largely absent from the performance of highly task-motivated individuals as it is believed these pilots were. These work decrement effects are also diminished when the work is of a varied rather than repetitive, routine nature. Furthermore, the overall work-period (about 13 hours) was not so long as to bring about

sufficient fatigue in these reasonably rested pilots to occasion much decrement, if strong task motivation is assumed, for there were many opportunities for rest during the workday.

However, of perhaps more concern than the possible trends with continuing work are the possible effects of task interference arising from the requirement to perform the tasks in close succession. This would be of special concern in the performance of C-8 and C-13, the pilots of which quickly passed from one phase to the other. Over the longer periods between phase performances of C-4 and C-9 the effects would diminish. Current understanding of transfer phenomena of such a nature predicts that the effect will be greater to the extent that responses to the same or similar stimuli in the two situations of concern are antagonistic. But, since the exact character of the intertask similarities can seldom be specified a priori, precise prediction of the effect is seldom possible and, therefore, again lacking comparative data, little beyond the minimal supposition that there may have been some intertask interference seems warranted. It may also be supposed on the grounds of greater possibility for antagonistic response that this effect would be most pronounced in the first phase performed at one station after having performed at another (i.e., on transferring stations). Of course, facilitative effects might also occur to the degree of positive intertask similarity in both stimuli and responses and they would be expected to parallel those of real-time operation with enhanced magnitude. On the other hand, such facilitative and interference effects as these are known to diminish with increased skill and with the frequency with which the tasks involved have been successively performed. All the crews had been required to shift from practice on one phase to another rather rapidly in the latter portion of training and, of course, C-8 and C-13 did so even more and were well-trained by usual standards. Thus, it is conjectured that although some additional facilitation and interference owing to the close succession in which the mission phases were performed may have occurred, it is not likely they were of sufficient magnitude to bias the main trend of the results. The effects of this nature, if existing at all, would be small changes up or down in parameter and phase reliabilities.

The multiple testing of the crews so that each crewmember performed in each position and the necessarily unique order in which this was accomplished represents another significant factor bearing on the results. However, rather than being forced to depend upon speculation about this factor the analytic method used permits direct assessment of most of the effects of concern. These have already been presented and provide, in general, a clear indication that the second- and third-tested individuals benefited from their immediate preexposure to the mission. As already noted, this factor might have interacted with the work decrement factor to produce some differential effects on C-8 and C-13 performance, but if the work decrement was slight these differential effects would also be slight.

The additional differences in order among the two sets of crews (C-4 and C-9 vs C-8 and C-13) deriving from the different way in which multiple testing was handled within the mission also might have produced some effects. Because of the great difference in performance level of the crews in the two study phases no deductions from the data concerning such effects seem possible. However, it may be surmised that if such effects occurred they were relatively slight and variable, and dependent upon the concerted influence of the continuing work and task transfer factors already discussed.

Another aspect of the testing procedure deserving special concern is the disturbance in the test schedule of C-8, necessitated by P-81's unplanned imminent departure on special assignment. Available information on this requirement was transmitted to P-81 in the initial portion of his (C-8's initial) mission. On completing the mission (near noon), and being uncertain as to exactly when he would have to leave, P-81 left the simulated crew station, donned street clothes, ate lunch in haste, and thus made all departure preparations. Consequently, he was not available the balance of the day to perform, in turn, as engineer and navigator in the subsequent test missions.

The consequences for the performance record of C-8 are, of course, difficult to gauge. But, it may at least be guessed that the probable effects, if any, were to degrade the performances of all the crewmembers. P-81, informed in midmission of a requirement of considerable personal importance, may well not have attended as fully as he might have to mission tasks. P-82 and P-83 were required to perform under an arrangement neither standard nor expected (though not unusual in their training experience) and may well have experienced minor confusion with perhaps some loss of motivation through disruption of crew integrity. However, if the performance of C-8 in the test mission was degraded in this manner it is suspected that the degradation was slight, primarily again because of the apparently still excellent task motivation and professional interest of the pilots. The disturbance in procedure represented mostly a nuisance to be accommodated for, not unusual in their experience.

As it turned out P-81 did not depart the first test day but was informed (upon being called off a commercial aircraft about to leave the passenger terminal for takeoff) that he should stay one day longer. This he did and was thus able to engage in retraining for one-half day before actually departing. Unfortunately, in having encountered still another change in plans, with what he interpreted as a possibly large personal sacrifice in his career and having not been able to recover his luggage, his motivation for the retraining suffered—to say the least. Exactly how much better he would have performed under the originally planned circumstances is uncertain but, judging from the present record it might have been considerably better. Therefore, it seems wise to place little value upon the retraining data of P-81, unless it is considered representative of performance under adverse conditions not typical of the conditions under which the others performed. These events also may have had consequences for the retraining performance of P-82 and P-83 for the same reasons mentioned above. However, it is surmised that such effects, if any, were slight—again for parallel reasons.

Finally, still with respect to the conduct of the test, at least two worthwhile issues can be raised regarding retraining. In the first place, what are the consequences for interpretation of the different amounts of training given the crews in the two study phases? That crews C-8 and C-13 were given much more retraining than C-4 and C-9 is, clearly, a matter of design attributable to the earlier experience with C-4 and C-9 and the resulting suspicion that they were not fully trained. Had C-4 and C-9 been given more retraining it would be possible to fully compare them with C-8 and C-13 and perhaps more fully assess the limitations in their original skill. However, judging from the data at hand, C-4 and C-9 probably would not have achieved their skill potential even if they had received the two additional days of training given the others. Therefore, it would still have been difficult to quantify, even after the fact, their original level of skill for the purpose

of relating this to skill retention and, in turn, increasing the value of the study. More retraining, unless sufficient for them to reach their skill potential, would not have rendered their data much more useful.

The second concern about retraining has to do with the way in which the retraining was given to C-8 and C-13. In particular, of what consequence is it that in most cases the individual performed four trials on a given phase in rapid sequence and then, being replaced by another pilot, went to another position or to a rest period? Certainly it might be supposed that some advantage might have been gained from frequent repetition of the same thing with fairly ample rest opportunities interspersed. On the other hand some disadvantage might have accrued from the disruption of the mission sequence and the need to warm up to performance of the phase scheduled. In general, however, it is supposed that any such effects were sufficiently slight and so variably distributed throughout retraining as to little influence the estimation of capability of primary concern in this study.

In fact, a similar view seems generally appropriate with respect to all the factors discussed relating to the conduct of the test. There is little question but what the various factors mentioned operated to drive the performance up or down on occasion. However, with the exception of P-81's retraining performance, it is unlikely that these effects were of any consequence. But, if there was any overall net effect of them it was probably performance-degrading rather than performance-enhancing.

Task Set of Test Personnel. One other feature of the data collection procedures, which is perhaps of greater significance than all the others, concerns the instructions given the test personnel about the nature of the test and what was expected of them. In daily circumstances as well as in formal laboratory experimentation the way in which a man views or defines his task (that is, his task goals) has a great influence on how he will perform and how he will divide his attention among the several aspects of the task. Consequently and because the tasks involved in this test were complex and multifaceted it is of great relevance for any interpretation of the data to ascertain the nature of the task goals held by the test personnel as they performed at various times. In particular, it is important to determine whether these goals may have shifted from training to test mission to retraining with resulting differential effects upon performance.

Now, for the fullest sort of information on how the pilots viewed their tasks, recourse would be necessary to some sort of interview and/or questionnaire technique intended to elucidate their thinking about them in depth. And to some extent this was accomplished with respect to task particulars, as reported by Grodsky et al (1966b), but it was not accomplished with respect to general goals. To have attempted to obtain information on general goals, except by perhaps the most subtle technique, would have engendered antagonism on the pilots' part. Their natural response would have been in the vein, "Well, of course, I did what you asked me to do." Furthermore, individuals are often unable to explicitly formulate the goals implicit in their activities.

Nevertheless, in the absence of formally taken direct expressions on general goals, it is possible to make some reasonably well-founded conjectures about them on the basis of the instructions given the test personnel and their informal comments to the test staff. On these grounds a rather clear distinc-

tion can be drawn between the probable approaches of pilots in C-8 and C-13 to their tasks in the test mission and in retraining versus their goals in training. A useful distinction also can be drawn between the handling of the two sets of crews in the two phases of the study.

Thus, the two crews of the earlier phase (C-4 and C-9) received rather general encouragement to perform well in training and subsequently in test and retraining. Specific goal criteria in flight control parameters were not identified for these crews as guides to their practice and performance, nor were they regularly provided precise information about their performance. No doubt they did operate with some awareness of what would be considered good and bad performance but these were never precisely defined.

In contrast, throughout their training and during the testing sequence C-8 and C-13 performed against known flight control goals for each phase, such as were listed in table V. (The actual criteria used included a few more than are represented in table V.) Pilots of these crews were also encouraged to do as well as possible, but it was thus additionally evident exactly what was considered good and bad performance. Furthermore, daily displaying to each pilot his previous day's performance by means of continuous graphs, with criteria indicated, served further to emphasize the urgency of meeting the criteria and to bring about close attention to rate of progress. In all likelihood these procedural differences in the way in which the two sets of crews were handled resulted in a more favorable learning situation for C-8 and C-13, at least up to the point of consistent capability to meet criteria in repeated trials. This effect, in addition to sheer differences in amount of training received, may be largely responsible for their greater skill at the end of training. By the end of training the pilots of C-8 and C-13 had achieved sufficient skill to regularly meet the criteria in most parameters, as table XXV, appendix III, shows. Clearly then, they had met the requirements of the original reliability study as defined for them and it was with this recall of their own former capability that they approached the skill retention test.

However, in the course of the initial briefing to the pilots the day before skill retention testing began, a strong attempt was made to redefine the task goals for C-8 and C-13 for the skill retention work. In these briefings the pilots of both crews were asked to seek the very best performance possible as consistently as possible, rather than to merely strive to meet the stated task criteria. By means of blackboard illustrations, they were shown how not only typical performance level but also variability in performance would be employed to compute reliability figures for their performances in the several parameters. And, they were told that these reliability figures would be taken as the primary indication of their capability.

These new instructions concerning task goals represented for the pilots a seemingly radical departure from previous instructions as well as from accepted practice in aircraft operation. As a result they naturally responded with some degree of confusion and requests for clarification. Clarification was provided until no more questions were forthcoming, but even then it seemed evident to the study staff that, although the pilots accepted the new task goals, they were not comfortable with them. Subsequent comments exchanged during the data collection, on review of test records, etc., tended to confirm this supposition and to suggest that the general attitude persisted throughout the study. Nevertheless, their comments also suggested that they were attempting to carry out instructions; i.e., to turn in the best possible performance as defined in the initial briefing.

To the extent the pilots actually adopted it, this reorientation to the mission tasks, as contrasted with the approach in training, probably has significant implications for the interpretation of the data. While earlier they merely had to meet the criteria, now they were to do that and, further, to exceed them as much and as consistently as possible. Certainly, with these new goals before them they would naturally tend to reexamine the several aspects of their performance in each phase, looking for ways of adjusting it by renewed or special attention to one aspect or another, particularly those in which they typically performed less well. They also would be expected to experiment with different operating strategies for accomplishing the more complex phases, such as braking and hover and earth entry, and one pilot did unintentionally confirm this expectation about his own performance.

The net effects of these actions on recorded performance would be a predictable gain in retraining so far as the criteria are concerned. However, the gain would not necessarily be regular or great for in seeking and shifting to new operational strategies some interim loss would be expected. Perfection of a new strategy would require at least a few trials. On the other hand, the effects of these same actions upon performance as interpreted by reference to former levels of estimated high reliability would be quite different. For, particularly the overall phase and mission reliabilities ($\hat{p}$ and $\hat{p}$) would be adversely affected by the great preponderance of compensatory shifts. To gain with respect to the criterion in a less adequately performed parameter would be to gain also with respect to former $p_{.95}$ level, to a probability >.950. But, to at the same time sacrifice some performance in another parameter, since the pilots were generally capable beyond the .999 level of meeting the criterion, often would result in no detected change in performance with respect to criterion, but always considerable loss below .950 with respect to former level. The nature of this differential effect will be examined in more detail later. For the present it merely serves to characterize the important implications of the change in task goals, which presumably occurred.

Analytic Procedures

Beyond the procedures used in data collection, there are still the specific analytic procedures used to summarize and organize the data obtained. The choice among these, too, can have a strong bearing upon the conclusions reached for they influence directly the nature of the analytic results obtained. This is particularly true of the present study which, being in some ways quite unlike traditional psychological research in goals and design, required the use of novel analytic procedures to render the data meaningful. The details of the analytic approach are given in appendix II, rather than being cumbersomely included within the body of this report, and no repetition of them will be given here. However, there are certain particulars which, because of their possibly great influence on the results obtained, should be explicitly recognized.

Sampling Decisions. Among these important analytic details are the decisions made regarding the sampling of data available. In the usual concern for sampling, the interest is in the set of test personnel employed and the manner in which this set is selected from the more inclusive population which they are supposed to represent. This usual concern is applicable to the present study as well and will be touched on in a following portion of this section. Here, the interest is in a different sampling problem involving the selection of a

limited portion of one individual's data to best represent that individual. Fundamentally, this problem occurs largely because in such data as these characteristic trends are expected; that is, it is anticipated that learning or forgetting will occur with consequent changes in performance. Thus, in the language of mathematical statistics, one is dealing with a nonstationary process. How then can one hope to arrive at an estimate of performance level at certain times from data series which must be assumed changing in one or more aspects? Beyond the nuances of this problem and more to the point of the present discussion, sampling for the purposes of this analysis was done in a certain way and this may have influenced the results in a predictable fashion.

Thus, in the first place, with but one exception, whenever there were sufficient data, the size of the performance sample taken was always four. Some retraining phases of C-4 and C-9 are the exceptions. This was done because four trials were estimated to be an optimum compromise between inaccuracies arising from the inclusion of trends and from error in parameter estimates from small samples. Obviously a larger sample size, say five or six, would have resulted in less estimation error if the data series was stable but not necessarily so if the series was still in the process of change. Whether four was the optimum is naturally debatable, but there is no doubt estimations deriving from a sample of four vary considerably, in accordance with the relatively greater weight accorded extreme values in the small sample and are occasionally in considerable error (so far as the mean of repeated such samples is concerned). It follows then that particular values obtained for p must be taken with qualification, especially when they deviate extremely, for they are overinfluenced by an occasional performance extreme. This would also be true to an even greater degree of p values taken jointly (i.e., $\hat{p}$) for the joint product is strongly influenced by inequalities among the elements involved. On the other hand p values taken cumulatively should represent quite accurate estimates, since any errors of estimation are presumably random and so tend to cancel out.

Besides sample size, one other aspect of the data sampling of special interest is the particular set of trials selected or the portion of the data taken. In having chosen a relatively small sample size and also elected to limit attention to only the last four training trials it is quite possible that, because of a low extreme in the trials selected, the individual's capability was underestimated. (In fact, correction in two such circumstances seemed necessary, as already described.) There is no reason to suppose that the last four trials were the individual's best, particularly if he was no longer learning rapidly. Thus, it might be suggested alternatively that the best four training trials should have been used as truly representing the individual's greatest attainment. Certainly this choice is a debatable one for taken at face value the best four trials indeed must be the best performance. On the contrary, to take the best four trials might be to overestimate the capability (mindful of the sample size), whereas use of the last four would result in random over- and underestimation and no overall bias. Also, from the practical point of view, finding the best four trials takes more work because more data must be analyzed.

Nevertheless, whatever the ultimate resolution of this issue may be, the last four trials were taken as the reference with the probable result that on some occasions the estimates derived over- or underestimated the "true" capability. A casual check by reviewing the results of an alternative weighted Z-score analysis performed on the same data did reveal a number of cases involving C-8 and C-13 in which a trial block other than the last apparently was superior.

But on closer examination of these by reference to comparable phase probabilities it was found that there are relatively few instances in which the actual superiority over the last trial block is considerable. For the C-8 pilots the shifts in $\bar{p}$ and $\dot{p}$ were as follows: P-81 in SDO, .919 to .998 and .757 to .995; P-82 in SDO, .953 to .987 and .860 to .960, and in LA, .806 to >.999; P-83 in TL1, .954 to .997 and .908 to .994. For the C-13 pilots the shifts in $\bar{p}$ and $\dot{p}$ were: P-131 in BH, .980 to >.999 and .920 to >.999; P-132 in SDO, .739 to .978 and .370 to .935, and in Dok, .939 to .980 and .819 to .940; P-133 in LA, .557 to .865, and in Dok, .942 to >.999 and .827 to >.999. The effects of substituting these new estimates of reliability in certain phases on estimated mission reliability may be seen in table XVIII, which shows both the old and new $\bar{\bar{p}}$, $\bar{p}$, and $\dot{p}$ values for individuals and crews. Obviously, the pilots' original capabilities (particularly P-132 and P-133) are estimated to be much higher when the best four-trial block performance is taken as the reference. But, it is still debatable whether this procedure gives a more valid estimate. Perhaps the most accurate value, if available, would be found to compromise between the two of table XVIII.

TABLE XVIII

Mission Reliability of C-8 and C-13 Per Criteria in Training as Estimated from the Last Four- and the Best Four-Trial Block *

Crew/Pilot		Measure $\bar{\bar{p}}$		$\bar{p}$		$\dot{p}$	
		last 4	best 4	last 4	best 4	last 4	best 4
C-8	1	.990	.999	.971	.997	.741	.974
	2	.973	.998	.961	.994	.683	.947
	3	.991	.996	.978	.988	.817	.894
	all	.984	.998	.970	.993	.747	.938
C-13	1	.967	.969	.946	.955	.582	.633
	2	.961	.993	.907	.983	.294	.854
	3	.932	.973	.907	.960	.361	.679
	all	.953	.978	.920	.966	.412	.722

* four-trial blocks taken as superior to the last four-trial blocks and substituted for them are only those which a weighted Z-score analysis suggested would be greater and for which $\bar{p}$ for that mission phase was subsequently found to be at least .008 greater

Furthermore, even if the best four-trial block in training is taken as a reference the consequences for the analysis of test mission capability to meet criteria are slight. For C-8 only SDO estimates are affected--P-81 shifting from .713 to .662 and .138 to <.001, and P-82 shifting from .980 to .996 and .939 to .989 in likelihood of meeting criteria ($\bar{p}$ and $\dot{p}$). The result in overall reliability for P-81 is a shift from .931 to .925, .827 to .812, and

.048 to <.001 in $\bar{\bar{p}}$, $\bar{p}$, and $\dot{p}$. The overall shifts for P-82 are .963 to .968, .920 to .929, and .395 to .431. In C-13, P-131 shifts from .294 to .250 in $\bar{p}$ on BH (no change in $\dot{p}$), with a resulting shift in $\bar{\bar{p}}$ from .899 to .894 ($\bar{p}$ and $\dot{p}$ unchanged). P-132 shifts ($\bar{p}$ and $\dot{p}$) from .893 to .999 and .695 to .997 in SDO and from .992 to .994 and .975 to .998 in Dok with changes in overall reliability ($\bar{\bar{p}}$, $\bar{p}$, and $\dot{p}$) from .982 to .995, .958 to .994 and .643 to .945. P-133 shifts from .716 to .885 in LPA and from .951 to >.999 and .853 to .999 in Dok with changes in overall reliability from .954 to .978, .927 to .962, and .481 to .697. Thus, as noted earlier, it appears that extreme values in the last four training trials are of concern only in the analysis of the test performance of P-132 and P-133. If these extreme values are considered spurious then capability to meet criteria in test must be considered somewhat underestimated for both P-132 and P-133—perhaps the most accurate estimate being some compromise value. Parallel reanalysis of test mission capability to meet former $p_{.95}$ levels, with the best four-trial blocks in training used in computing those levels, results in probabilities that are never greater and are generally less than those originally obtained. Although these reductions in probability are small and of little consequence, again the probabilities for P-132 and P-133 are found most affected. The values for P-132 shift from .871 to .831 in $\bar{\bar{p}}$, from .731 to .627 in $\bar{p}$, and from .004 to <.001 in $\dot{p}$. Those for P-133 shift from .805 to .758, from .570 to .491, and from .001 to <.001. Accordingly, the capabilities of P-132 and P-133 to meet former $p_{.95}$ levels may be considered slightly overestimated to the extent that extreme values recorded for them at the end of training are spurious.

Comparison of best retraining performance with best training performance still shows a gain in capability to meet criteria for four of the six pilots. By this comparison P-81 never did quite as well in retraining as in training (.997 vs .998 in $\bar{\bar{p}}$ and .978 vs .995 in $\dot{p}$), probably for the reasons already discussed. Also, P-133 did not demonstrate his original capability to meet criteria (.991 vs .999 in $\bar{\bar{p}}$, .964 vs .996 in $\bar{p}$, and .789 vs .976 in $\dot{p}$). The cumulative probabilities for crews still indicate a net gain in capability to meet criteria in retraining. The parallel analysis of greatest demonstrated capability in retraining to meet former $p_{.95}$ levels, in making reference to occasionally superior phase performance, can never result in probabilities greater than those based on the last trial-block reference.

Thus, it is concluded that the use of the last four training trials as a basic reference had little affect upon the main outcomes of the analysis of C-8 and C-13 performance. It apparently resulted in underestimation of original capability to meet criteria, of much consequence only for P-132 and P-133. In turn, the capabilities of P-132 and P-133 for meeting the criteria in the test mission appear somewhat underestimated, whereas their capabilities to meet former $p_{.95}$ levels appear slightly overestimated. However, the critical estimations on test performance of P-81 and P-131, on which the main conclusions must rest, are virtually unaffected. The outcomes of comparisons involving retraining also remain essentially unchanged, even though use of the last four training trials may have introduced a slight overestimation of capability in retraining.

A similar but still more complicated dilemma exists with the selection of a four-trial block to best represent capability in retraining. Rather obviously the first trial block would not be the best if relearning or new learning is assumed. But, the last one might not be the best either, particularly if there

is some evidence of or basis for assuming contradictory trend effects from the presence of some factor other than learning. Unfortunately, the presence of just such an effect does seem indicated by the retraining data for C-8 and C-13, and staff observation suggests a continuing degeneration in the task motivation of the test personnel after the first day of retraining as the accountable factor. Therefore it seems most reasonable in this instance to take the best four trials as most representative of individual capability—the estimates derived from them being inflated by sampling bias, but deflated by reduced motivation in roughly a compensatory fashion. It is a matter of conjecture just how truly representative the estimates so obtained are, but it is believed they are probably reasonably close (say within .01 in $\bar{p}$).

However, the choice of the best four trials does not exhaust the ramifications of the sampling problem with respect to retraining data. It was still necessary to decide whether the best four trials with respect to $\bar{p}$ or $\dot{p}$ should be used and whether these should be with respect to criteria ($\bar{p}_c$ or $\dot{p}_c$) or whether they should be with respect to the former attainment in training ($\bar{p}_{.95}$ or $\dot{p}_{.95}$). For the main analysis of results, comparing the several levels of performance exhibited throughout the experiment (see table XIV), this was resolved by taking $\dot{p}_c$ for the criterion-referenced analysis and $\dot{p}_{.95}$ for the former level-referenced analysis. The $\dot{p}$ was considered superior to $\bar{p}$ as best or more sensitively reflecting the overall capability in any phase.

This solution does bring about an analytic dilemma stemming directly from the fact that the best four-trial block indicated by $\dot{p}_{.95}$ performance is more than a few times not the best trial block indicated by $\dot{p}_c$ performance. In other words, two somewhat different sets of retraining trials are used to represent retraining capability. This circumstance might surely have implications for the comparison of gains/losses in retraining over original levels as derived from criterion-referenced versus former level-referenced probabilities.

Such a comparison of the performances of C-8 and C-13 is summarized in table XIX. Clearly, as stated in the final portion of the presentation of results, so far as the criteria go these crews gained, but so far as former $p_{.95}$ level is concerned they lost in retraining over end-of-training capability. The question which naturally follows is whether this contradictory result is to be accepted as plausible and solely the result of the shifts in task set already described, or whether it may not also be partly the consequence of the selection of best four-trial block in retraining. Specifically, what would be the alternative results if $\dot{p}_{.95}$ were computed on the basis of the best criterion-referenced block when estimated capability to meet criteria is maximal (becoming $\dot{p}_{.95/c}$)? Similarly, what would happen if $\dot{p}_c$ were computed on the basis of the best $\dot{r}_{.95}$ level-referenced block when capability to achieve former $p_{.95}$ levels in individual parameters is maximal (becoming $\dot{p}_{c/.95}$)?

Several deductions about the relative orders of magnitude of the alternative $\dot{p}$ values and the cross-comparison of them can be made. In the first place $\dot{p}_{c/c}$ must be greater than $\dot{p}_{c/.95}$, because $\dot{p}_{c/.95}$ is based on at least some smaller p values. Since $\dot{p}_{c/c}$ in retraining is known to represent a gain over training, $\dot{p}_{c/.95}$ must be found to show either less gain, no difference, or possibly a loss. Similarly $p_{.95/.95}$ must be greater than $p_{.95/c}$ because the latter is based on some smaller p values. Further, since $p_{.95/.95}$ has been found to represent a loss, $p_{.95/c}$ must be found to represent still greater loss. Thus, in the comparison of gain/loss of $p_{c/.95}$ and $p_{.95/.95}$ the differences must

TABLE XIX

Change in Mission Reliability of C-8 and C-13 from Training to Best Four-Trial Block in Retraining *

Crew/Measure			Pilot 1	Pilot 2	Pilot 3	All (mean)	p Expected
C-8	$\bar{p}$	c	.011 (1.1)	.008 (0.8)	.002 (0.2)	.007 (0.7)	vario
		.95	-.306 (-32.2)	-.087 (-9.2)	-.121 (-12.7)	-.171 (-18.0)	.950
	$\bar{p}$	c	.033 (3.4)	.026 (2.7)	.005 (0.5)	.021 (2.2)	vario
		.95	-.583 (-67.9)	-.234 (-27.3)	-.227 (-26.5)	-.348 (-40.6)	.858
	$\dot{p}$	c	.221 (29.2)	.152 (17.9)	.028 (3.1)	.133 (16.7)	vario
		.95	-.393 (-99.0)	-.372 (-93.7)	-.392 (-98.7)	-.386 (-97.1)	.397
C-13	$\bar{p}$	c	.014 (1.4)	.050 (5.3)	.002 (0.2)	.022 (2.3)	vario
		.95	-.058 (-6.1)	.019 (2.0)	.001 (0.1)	-.013 (-1.3)	.950
	$\bar{p}$	c	.044 (4.6)	.126 (14.6)	-.003 (-0.3)	.056 (6.3)	vario
		.95	-.172 (-20.0)	.044 (5.1)	-.021 (-2.4)	-.050 (-5.8)	.858
	$\dot{p}$	c	.246 (33.3)	.636 (211.3)	-.018 (-2.2)	.288 (80.8)	vario
		.95	-.329 (-82.9)	.124 (31.2)	-.109 (-27.5)	-.105 (-26.4)	.397

* percentage gain or loss indicated in parentheses

TABLE XX

Change in Mission Reliability of C-8 and C-13 from Training to Best Four-Trial Block in Retraining per Alternate Reference Measures *

Crew/Measure			Pilot 1	Pilot 2	Pilot 3	All (mean)	p Expected
C-8	$\bar{p}$	c/.95	-.090 (-9.1)	.004 (0.4)	-.046 (-4.6)	-.044 (-4.4)	vario
		.95/c	-.330 (-34.7)	-.184 (-19.4)	-.100 (-10.5)	-.205 (-21.5)	.950
	$\bar{p}$	c/.95	-.203 (-21.1)	.009 (0.9)	-.140 (-14.2)	-.111 (-11.5)	vario
		.95/c	-.591 (-68.9)	-.447 (-52.1)	-.262 (-30.5)	-.433 (-50.5)	.858
	$\dot{p}$	c/.95	-.477 (-63.0)	.047 (5.5)	-.642 (-71.3)	-.357 (-42.9)	vario
		.95/c	-.395 (-99.5)	-.397 (-100.)	-.396 (-99.7)	-.396 (-99.7)	.397
C-13	$\bar{p}$	c/.95	-.007 (-0.7)	.047 (5.0)	.002 (0.2)	.014 (1.5)	vario
		.95/c	-.077 (-8.1)	-.062 (-6.5)	.019 (2.0)	-.040 (-4.2)	.950
	$\bar{p}$	c/.95	-.591 (-68.9)	.115 (13.3)	-.262 (-30.5)	.024 (2.9)	vario
		.95/c	-.227 (-26.5)	-.128 (-14.9)	.028 (3.3)	-.109 (-12.7)	.858
	$\dot{p}$	c/.95	-.251 (-34.0)	.575 (191.0)	-.046 (-5.7)	.093 (50.4)	vario
		.95/c	-.370 (-93.2)	-.326 (-82.1)	.011 (2.8)	-.228 (-57.5)	.397

* percentage gain or loss indicated in parentheses

be less than that apparent in table XVII, while in the comparison of gain/loss of $p_{.95/c}$ and $p_{o/c}$ the differences must be still greater. Furthermore, the probabilities obtained should be ordered from most gain to most loss as follows:

$$p_{o/c} > p_{o/.95} ? p_{.95/.95} > p_{.95/c}$$

These deductions may be confirmed by cross-comparing the tabulations of p computations based on $p_{.95/c}$ and $p_{o/.95}$ given in table XX with those of table XIX. Which, then, of the three possible pairs is the most valid and useful one for representing retraining capability or should all, perhaps, be used in concert to provide a fuller understanding of the study results?

The latter view seems unquestionably the more appropriate one. Clearly $p_{o/c}$ comparisons, as representative of capability to meet the task goals underlying all performance, are of primary concern. The retraining over training gain in them must be taken at face value as indicative of genuine improvement. In addition, that the crewmembers showed large losses as in the $p_{.95/c}$ comparisons, while simultaneously gaining with respect to criteria, suggests a definite change in approach to the task. That this change cannot be considered to any great extent an artifact of the retraining trial sample used is demonstrated by the similarly low $p_{.95/.95}$ values. Evidently, although the problem of sampling in retraining introduces complexities in analysis it does not pose particular difficulty for interpretation. Certainly the differences in results concerning retraining would be to only a slight extent, if any, the consequences of the sampling used.

The Statistical Model and Simple Indicants. Tacitly assumed in the foregoing discussion of the sampling problem is that the sample data are to be interpreted in terms of some statistical model. In this instance, the model adopted is the very commonly used normal density function. Thus, in using this model it is being assumed that the mean, standard deviation, and other parameters of a hypothetical sampling distribution of the data samples under examination approximate those of a normal distribution. If this assumption is acceptable then the interpretation of the data sample in terms of the theoretical normal distribution leads to no systematic bias. On the other hand, if it can be shown, or there is a reasonable basis for assuming that the sample data depart from the normal in certain ways, then the possible biases introduced must be considered. There are at least three specific questions concerning use of the normal distribution which should be discussed.

First, although many human performance measures have been found to approximate normal distributions this is not invariably true and the question may properly be raised as to whether the measures used in this study did distribute normally. If performance had at any time stabilized (on some acceptable criterion) over an appreciable number of trials a rough answer to the question might have been obtained by the use of tests for normality. But, because trends were likely and a four-trial sample was chosen as the analytic base standard tests for normality were not feasible. Accordingly, the question must be answered simply (if at all) on the basis of an examination of the data and awareness of measurement factors commonly resulting in nonnormal distributions.

Review of the data did not, however, disclose evidence of sufficiently frequent departures from normality as to give pause for serious concern. There were, of course, a few instances in which the last four training trials did not seem appropriately variable or seemed skewed, but this is to be expected with so small a sample. Certainly the few instances can hardly be taken as proof that the normal distribution was not generally applicable. Even the critical velocity cutoff error, treated as it was without regard to sign, did not clearly exhibit the degree of asymmetry which might be expected. Apparently, that there was almost always some such error and that it tended to be biased the same direction from zero resulted in this measure bearing a reasonable semblance of a normal distribution only moderately truncated by the zero bound. The other error measures (displacement and displacement rate) which also might be supposed to present a similar problem apparently were sufficiently large as not to present serious truncation or asymmetry either. On the other hand, it is not being contended that sampling distributions of these measures (and perhaps certain of the others) would be normal—only that they would be sufficiently close to normal that the application of the normal model is feasible.

Apart from these possible gross departures from the normal as a model, there is the obvious arbitrary use of it to arrive at probabilities, as simple indicants of performance capability in the measured flight control parameters. The chosen sample size was indeed small—much smaller than is commonly interpreted by reference to the values of the normal density function. With such small samples of an otherwise normal variate, the Student t distribution is the traditionally-recommended model for a closer approximation to the true probabilities. And, unquestionably, it would have provided more exact values for these data as well.

However, the use of the normal rather than the t distribution seemed justified in this instance on two bases. These are the ready availability of extensive normal probability tables showing values to the nearest thousandth and the nature of the interest in the data. As has already been noted and will be made still more evident, the absolute levels of the probabilities reported would, in any case, have to be taken with considerable qualification and, furthermore, the primary interest in the study data is in performance changes or differences rather than in absolute level of reliability. Therefore, the inconvenience of using the t distribution did not seem warranted. Inaccuracies in p arising from this procedure would tend to run parallel in any two sets of performance and so should not significantly affect the outcome. But, it is true that in having been obtained from normal rather than t tables, the values reported must be considered somewhat inaccurate.

The nature and extent of these inaccuracies may be judged readily on the basis of the comparison figures given in table XXI. From this table, indicating the performance of P-131 in terms of both the normal (z) and the t distributions, it seems clear that the inaccuracies introduced by using z are modest and follow a consistent pattern. By using z the probabilities of meeting the hypothetical criteria are overestimated, but in lesser amounts as one progresses from training, to test, and then retraining. In the worst case (training) $\bar{p}$ is enhanced only .020 and even $\dot{p}$ is enhanced only .228, which is modest considering its extreme sensitivity. Such overestimation with respect to criteria is easily understood to result from the fact that at this pilot's high level of skill the probability associated with a given z is nearly always greater than the probability associated with t of the same value. Because fixed criteria are references, the z and t values in each instance are the same.

TABLE XXI

Performance of P-131 in Training, in the Test Mission and in the Best Four-Trial Block in Retraining per Normal and χ^2 Distributions

Phase/ Measure		Trng/c		Test /c		Test/p.95		Retrng/c		Retrng/p.95	
		N	χ^2	N	χ^2	N	χ^2	N	χ^2	N	χ^2
TLI	$\bar{p}$	—†	.992	—	.999	—	.998				
	$\hat{p}$	—	.985	—	.999	—	.995				
THM		—	.996	—	.994	.591	.629	—	—	.999	.997
		—	.989	—	.983	<.001	.024	—	.999	.997	.992
LOI		.999	.987	.894	.989	.984	.973	—	.999	.886	.919
		.999	.974	.789	.978	.968	.946	—	.998	.771	.838
NDO		.932	.900	.960	.923	.853	.915	—	.997	—	.999
		.804	.721	.880	.782	.600	.761	—	.992	—	.997
EM		.980	.925	.294	.297	.285	.359	—	.997	.883	.911
		.920	.722	<.001	<.001	<.001	<.001	—	.989	.534	.651
LPA		.999	.976	.903	.857	.382	.645				
		.999	.976	.903	.857	.382	.645				
Rcnd		.789	.760	.936	.887	.991	.973				
		.789	.760	.936	.887	.991	.973				
Rck		—	.997	—	.997	.653	.694	—	—	.788	.799
		—	.990	—	.990	.026	.144	—	—	.366	.409
EM		—	.996	—	.995	.755	.836	.995	.975	.796	.815
		—	.988	—	.984	.301	.540	.985	.925	.451	.494
All	$\bar{p}$	.967	.947	.899	.882	.722	.780				
	$\hat{p}$	.946	.899	.858	.829	.474	.559				
	p	.582	.354	<.001	<.001	<.001	<.001				
Rotng	$\bar{p}$	.985	.967	.876	.866	.687	.734	.999	.995	.892	.907
	$\hat{p}$	.954	.897	.813	.786	.316	.403	.998	.984	.686	.730
	p	.739	.490	<.001	<.001	<.001	<.001	.985	.905	.068	.109

† dash indicates value >.9995

On the other hand, by using z the probabilities of meeting former $p_{.95}$ levels are underestimated, apparently again in lesser amounts as one progresses from test to retraining. Once again the differences are not disturbingly great, with $\bar{\bar{p}}$ depreciated only .058 in the worst case and the other indicants ($\bar{p}$ and $\dot{p}$) relatively less affected, in view of their sensitivity. This underestimation results from the fact that in computing the probabilities with respect to former $p_{.95}$ levels, the reference level itself changes with the use of z, rather than t, so that it is generally less stringent than that obtained by the use of t. With the unchanged standard deviation estimate as the divisor the calculated value of z is necessarily less—evidently sufficiently less (with the frequently modest probabilities found in this case) to reduce the more inclusive probabilities below that obtained by using t.

Thus, in having used the normal rather than the t distribution as a model criterion-referenced probabilities are slightly overestimated and former $p_{.95}$ level-referenced probabilities are slightly underestimated. However, as was supposed, the differences of interest are not particularly affected by these inaccuracies. When calculated by t the losses in capability of P-131 to meet criteria in test compared with end-of-training capability are 6.9, 7.8, and 100% for $\bar{\bar{p}}$, $\bar{p}$ and $\dot{p}$, respectively. These compare favorably with the parallel percentages of 7.0, 9.3 and 100 given in table IX for the z analysis. The losses (by t) in capability in test to meet the former $p_{.95}$ level of 17.9, 36.7 and 100% are also consistent with the parallel percentages of 24.0, 46.3 and 100% given in table XI. The gains (by t) in maximum retraining capability to meet the criteria of 2.9, 9.7 and 84.7% are somewhat greater than those of 1.4, 4.6, and 33.3 reported in table XIX. Also, the losses (by t) in maximum retraining capability to match former $p_{.95}$ levels of 4.5, 14.9, and 72.5 are somewhat less than the parallel percentages of 6.1, 20.0, and 82.9 shown in table XVIII. But, the primary effects remain the same as previously given in all instances. Evidently, on the basis of the data from P-131, the discrepancy in findings on retraining (p_c vs $p_{.95}$) is not likely to any significant extent the consequence of having used the normal rather than the t distribution as a model.

One final concern regarding the computation of the basic probabilities has to do with the calculation of those for the test mission. Obviously only one data point exists for each parameter and to have arrived at a probability estimate on the basis of just one measure may seem to be the epitome of statistical bootstrapping—and perhaps it is. However, the fundamental operation of calculating a probability value to represent this single measure in reference to some criterion (given or derived from the individual's own performance) is no different from the commonplace acceptance of the mean of the raw scores as providing the best available estimate of performances which would occur under those circumstances in repeated testing. Since the mean is, in a symmetrical distribution, also the $p_{.50}$ value the logic is precisely the same. It is true that the sampling distribution of $p_{.95}$ is greater than that for $p_{.50}$ for any given n but this is a difference in accuracy, not in kind. Thus, it seems the inherent nature of the $p_{.95}$ value does not introduce any difficulty. What, then, is the difficulty with it?

Instead, the problem and, in turn, possible source of bias arises because in necessarily limiting attention to one measurement (to avoid contamination from learning) the opportunity to make a direct estimation of variability is sacrificed. Without such an estimate the meaning (range) of the $p_{.50}$ value

(or mean) cannot be known and, the $p_{.95}$ value or criterion-referenced probability cannot even be calculated. Thus, it is important for the interpretation of the data that some such estimate be available. Fortunately in this situation, as in many others, a sufficiency of information on performance exists to form a basis for estimating the variability which would have been observed on repeated testing. The important question is merely, What bias or other effects did the use of a borrowed variance estimate in calculating p introduce? Or, put the other way, Did the use of a borrowed variance estimate result in an acceptably accurate estimate of the probabilities in question?

In this instance, rather than to complicate the analysis still further by special calculations such as might be used in deriving the most accurate estimates possible, the variances observed in the last four training trials were used in obtaining the test mission probabilities. If the test mission performance had been exactly comparable to end-of-training performance it could be argued that any errors arising from this procedure would be randomly distributed and so self-canceling. But, the test mission scores are, in fact, somewhat poorer than end-of-training performance means. This, together with the tendency of variability to change in the same direction as the mean, implies that the variances used in calculating probabilities for the test mission performance were more frequently underestimations of the variances which would have been obtained by multiple testing (if that were possible) than they were overestimates. Since the standard scores were obtained by dividing the difference between test mission and criterion or $p_{.95}$ level reference values by the estimated standard deviation (s), the net result is that there is a tendency for probabilities derived from scores better than the reference values to be overestimates and those from scores worse than the reference values to be underestimates. In general, then, there resulted from the use of borrowed variance a bias in the direction of overestimating the test mission capability. However, because the degradation in performance was typically small it may be supposed that the "true" variance would not be greatly different from that used and, in turn, that the magnitude of this bias is slight. It is estimated not to have exceeded .02 although any such exact estimate is subject to doubt.

The Reliability Model and Complex Indicants. Finally, there are many additional matters that could be discussed concerning the way in which overall performance was indicated. Of course, even the use of a probalistic or reliability model for arriving at overall indicants may be considered dubious because it is at variance with psychological traditions. However, assuming agreement on that general question, it is still appropriate to inquire whether the particular way in which performance was represented probabilistically in this analysis resulted in any biases to be taken account of in interpreting the results. Two rather evident possibilities of this sort seem worthy of special comment.

The first of these concerns the attenuation of $\bar{p}$ with compensatory change in the contributing elements and with $\bar{p}$ unchanged. This phenomenon has already been described briefly in illustrating the concern for task set and the apparent results of it are exemplified by the relative gains and losses in retraining as depicted by tables XIX and XX. But, for a more complete understanding it may be helpful to consider this simplified example.

Suppose that overall performance on a two-parameter phase is of interest and that the probabilities of meeting the former $p_{.95}$ level in both parameters and on either parameter are considered to satisfy this interest. By definition the individual approaches the task with an expectation based on past performance of meeting the former level in a parameter 95% of the time and meeting it in both 90.25% of the time (.950 x .950). Now, what may actually happen to his performance relative to his former level? Clearly, he may meet it on both, fail it on both, or fail it on one but not on the other.

Now, if the estimate of the $p_{.95}$ level, accepted as a reference, is subject to random variation then it should be expected that in the long run successes (S) and failures (F) would be equally distributed across the two parameters. The result would be that the $\bar{p}$ would always be .950, as can be illustrated by taking all possibilities with an equal gain or loss of, say, .050, being assumed. The resulting $\bar{p}$ values are 1.000, .950, .950, and .900 for the combinations SS, SF, FS, and FF, respectively. The average for the $\bar{p}$s in the total set, which if ps are varying randomly must occur with equal frequency, is still .950. (This is true of any collection of such sets also—i.e., for $\bar{\bar{p}}$.)

In a similar manner, the $\dot{p}$ values resulting from equal gains and losses of .050 may be determined and contrasted with the expected value of .9025. Again assuming for the sake of convenience equal gains and losses of .050 they are 1.000, .900, .900, and .810 for SS, SF, FS, and FF, respectively. The average for the $\dot{p}$s in the total set (which again if ps are varying randomly must occur with equal frequency) is indeed .9025, as would be expected when no change in performance has occurred. But, note the difference between this circumstance and that of $\bar{p}$. Whereas compensatory gains and losses among p values result in no deviation from the expected $\bar{p}$ (the sum of 1.000 and .900 divided by 2 still equals .950) they do result in a deviation from the expected $\dot{p}$ (.900 versus .9025, in this example). In addition, the mutual gain brings about a change in $\dot{p}$ which is greater than that brought about by the equivalent mutual loss. And, furthermore, these attenuations of $\dot{p}$ are compounded still more in the calculation of $\dot{\bar{p}}$ for a collection of such sets.

From this example it may be concluded that if any factor (such as learning) enters into the test performance to disturb the random frequencies and magnitudes of gains and losses occurring in the several parameters it will be duly reflected in the complex indicants $\bar{P}$, $\bar{\bar{p}}$, $\dot{p}$, and $\dot{\bar{p}}$. $\bar{P}$ and $\bar{\bar{p}}$ will be shifted up or down equally by the same amount of mutual gain or loss in p and will be unaffected by compensatory shifts. However, $\dot{p}$ and $\dot{\bar{p}}$ will be disproportionately affected such that they will be increased by mutual gains in p to a greater extent than they will be degraded by comparable mutual losses of the same extent, and exactly compensatory gains and losses will degrade them. These disproportionalities vary directly with the amount of shift starting with quite small deviations, as in this example. Clearly, when compared with $\bar{p}$ and $\bar{\bar{p}}$, $\dot{p}$ and $\dot{\bar{p}}$ place a premium upon the individual maintaining his relative performance on the several measured aspects of the task and reflect stringently on any tendency to gain in some and lose in other aspects, in tradeoff fashion.

Now, having recognized the disproportionalities which may occur in $\dot{p}$ (and $\dot{\bar{p}}$ in turn) it might easily be supposed that $\dot{p}$ is not a useful indicant of performance in a complex (multimeasure) task. However, that is not the contention here nor is it a necessary conclusion. For example, it seems entirely reasonable to argue that $\dot{p}$, as a complex indicant, has exactly the properties that an indicant

useful in reliability analyses should have. On the face of it, at least, there is no doubt that it duly reflects the probability of success in all task aspects—surely a useful concept—hence, that is not the problem which it occasions. Instead the problem which it poses seems merely to be an interpretive one of relating the possible variations in it to traditional views on human performance. There is a strong convention of thinking about multiple performances in terms of average values (arithmetic means, etc.) and, of course, $\bar{p}$ follows this convention. Accordingly, one expects equal gains and losses to be entirely self-cancelling. But, there is no compelling reason why they should be and, indeed, there may be interpretive value in their not being so. For, in a broad view of the human performance, is not inconsistency in dealing with the task also a reflection of inadequacy and unreliability? If this is so, then the use of $\dot{p}$ is entirely justified. Nevertheless, its use does require special care in interpretation, as it has been the primary intent of the present elaboration to show.[3]

Thus, briefly returning to the analysis of retraining performance, the characteristics of $\dot{p}$ and the insensitivity of the criterion-reference analysis to certain losses taken in conjunction with the likely task set of the pilots (of C-8 and C-13) easily could result in the contradictory findings obtained. As capable as the pilots were in meeting the criteria (p much greater than .999 in many parameters) the gains with respect to parameters in which they were least capable would bring overall gains whereas the parallel losses in the parameters of greatest capability would often not be detected in the analysis with respect to criteria. In contrast, the analysis with respect to former level, being sensitive to shifts in both directions and harshly reflecting any compensatory shifts, would bring overall losses in $\dot{p}$ and the more inclusive indicants based on it. Only $\bar{p}$ and $\bar{\bar{p}}$ would directly represent the magnitude of the overall shift unaffected by the way in which the shift occurred. That $\bar{\bar{p}}_{.95}$ did indicate some loss suggests that gain in one parameter cost disproportionately more in another.

Support for this conjecture as, at least, partly explaining the contradictory retraining results may be derived from an analysis of the relative order of capability on the several parameters of a given phase in training and in retraining. If the supposition of a shift in task goals by the pilots of C-8 and C-13 is true, then their retraining performance evaluated by reference to the former $p_{.95}$ level ought to be negatively correlated with their end of-training capability to meet criteria. But, retraining capability to meet cri' a contrasted with end-of-training capability would be negatively related only if shifts larger than interparameter differences occurred: otherwise with small shifts the order of performance would be maintained and the correlation would be positive.

---

3 Actually, the arithmetic effects of compensatory shifts on these data apparently are not as great as might be supposed. If the $p_{.95}$ values for P-131 in his best four-trial block in retraining are adjusted to remove all effects of compensatory shift (by equalizing parameter values within each phase at the level of $\bar{p}$ for the phase) the changes obtained in $\bar{\bar{p}}$ and $\dot{p}$ are modest. Thus, $\bar{\bar{p}}$ increases from .686 to .730 and $\dot{p}$ increases from .068 to .117. Furthermore, the losses in capability relative to end-of-training skill are only somewhat reduced. In $\bar{\bar{p}}$ the loss of .172 (20.0%) becomes .128 (14.9%) and in $\dot{p}$ the loss of .329 (82.9%) becomes .212 (53.4%). $\bar{\bar{P}}$ values are not affected, of course.

The results of just such an analysis, using $\dot{p}$ values rather than p for greater discriminability, is tabulated by pilot and retraining reference in table XXII. In all of the 33 cases represented by this table there are only 9 in which the critical order of performance per the $p_{.95}$ levels (with $\dot{p}_{.95/.95}$ the reference) is not negatively related to training performance in the predicted fashion. In view of the insensitivity of the analysis to partial tradeoffs (for example, compensatory shifts in only two of three parameters) this proportion of variant cases does not seem large. Hence, it is concluded that the presumed task set effects did occur, although not entirely consistently either among individuals or among phases. Among the pilots, 82, 131, and 132 showed the effect more than the others: among phases, the effect is most pronounced in separation and deorbit, braking and hover, and docking and least pronounced (or nonexistent) in transposition.

Furthermore, if transposition is discounted, a loss in capability to meet former levels ($\dot{p}_{.95/.95}$) is seen in 16 of the remaining 21 cases in which an inverse relation in order of performance is noted. The overall average shift for all cases (including transposition) in which inverse ordering occurred represents a loss, the average shifts being -.448, -.226, -.255, -.234, +.042, and +.024 for pilots of C-8 and C-13, respectively. (The slight gains of P-132 and P-133 are attributable to the influence of extreme values at the end of training, which has already been mentioned.) The comparable average shifts for cases in which a positive relationship was found consistently represent less loss or more gain, except that for P-81 whose retraining data are discounted. These shifts are -.806, .140, -.086, .143, .050, and .097 for C-8 and C-13 pilots, respectively.

Although the specific gains and losses in phases do not correspond perfectly with direct and inverse performance orders, this is to be expected. A compensatory shift would not necessarily imply loss (general gain in all parameters could occur) just as continuing the same performance order would not always result in gain (a general loss could occur). Thus, it is not contended that all gains and losses are accounted for by task set shifts. Rather it is suggested that set shifts resulted merely in a trend toward degradation in phase performance evaluated by reference to the former $p_{.95}$ level in individual parameters. This, admittedly weaker, relationship does appear verified.

The other special concern relating to the way in which overall performance is represented has to do with the possible relationships among the several measured aspects of task performance. In psychological analysis of day-to-day behavioral complexities it is generally understood and expected that some of the measured or measurable aspects of the performance of interest may be in some kind of dependent relationship. This dependency may result from a contingency evident in the elements of the task, or from some common skill requirement, or from some other source—exactly how not necessarily being of initial significance for the measurement operations. The main problem is the determination of what the relationship among the measures is. With this information available determination of the basis for and detailed nature of the relationships may then be undertaken.

In keeping with this accepted view on complex tasks it is entirely appropriate to inquire about the possible interdependencies among task aspects in the present study. If there were task dependencies, the next and also appropriate question to be asked is, How are these dependencies reflected in the

TABLE XXII

Relationship Between Order of Parameter Performance in Training and Best Four-Trial Block in Retraining (Kendall's τ)

			Pilot						
Reference/ Measure			81	82	83	131	132	133	All (mean)
THN	p	c/c	ND	1.000	-.333	.333	.333	1.000	.467*
		c/95	ND	1.000	-.333	.333	.333	1.000	.467*
		95/c	ND	.333	-.333	-.333	.333	1.000	.250*
		95/95	ND	.333	-.333	-.333	.333	-.333	-.067*
LOI			ND	-	+	+	+	+	+ *
			ND	-	+	+	+	+	+ *
			ND	-	-	-	-	+	- *
			ND	-	-	-	+	+	- *
SDO			-1.000	.333	1.000	.333	-1.000	-.333	-.111
			-1.000	.333	1.000	.333	-.333	-.333	.000
			-1.000	-1.000	-.333	.333	-1.000	-.333	-.556
			-1.000	-1.000	-.333	.333	-.333	-.333	-.444
BH			.333	1.000	.000	.667	.667	.333	.500
			.333	1.000	.000	1.000	.667	.333	.556
			-.333	-.333	-.667	-.667	-.667	.000	-.445
			-.333	.000	-.667	-1.000	-.667	.000	-.445
Dok			ND	.333	-.333	.333	-.333	-.333	-.067*
			ND	1.000	1.000	-1.000	-.333	-.333	.067*
			ND	-1.000	-1.000	-.333	-1.000	-.333	-.733*
			ND	-1.000	1.000	-1.000	-1.000	-.333	-.467*
EE			1.000	1.000	.333	-1.000	-1.000	.333	.111
			1.000	.333	.333	-1.000	-.333	.333	.111
			-1.000	.333	-.333	-.333	-1.000	-.333	-.444
			1.000	-.333	-.333	-.333	-.333	-.333	-.111
All	p	c/c	.111*	.733	.133	.133	-.267	.200	.180
		c/95	.111*	.733	.400	-.067	.000	.200	.240
		95/c	-.778*	-.333	-.533	-.267	-.667	.000	-.386
		95/95	-.111*	-.400	-.133	-.467	-.400	-.266	-.307

* based on reduced data
ND not done by this pilot

reliability model by means of which overall capability is indicated? Now, it is already evident that the simple probability model used in the present analysis to obtain $\bar{p}$ and $\dot{p}$ and the more inclusive indicants based on them, frankly provides no means of introducing task dependencies. Each parameter p-value is taken on its own merits as in no way dependent on any other and if all are independent, as is assumed, $\bar{p}$ and $\dot{p}$ are accurate within the limits of measurement error. On the other hand, performance in certain parameters may be related in actuality, with the result that to treat them as independent is to introduce a bias. Hence, the question is, To what extent were performances in the several parameters related and how, if any, are the results biased thereby? Once again it is not possible to give a complete and entirely satisfactory answer and it is necessary, as with previous issues, to fall back upon analysis or personal judgment in formulating an opinion.

By the way of exploring the seriousness of this issue an attempt was made to quantify the interdependencies among task elements as represented by the performance record. Since there is no reason to expect that whatever relationships existing would be the same throughout the acquisition of skill in the tasks--in fact, it is more reasonable for them to differ--it was necessary first to choose the reference sample for the analysis. The last four training trials had already been used as the basic reference in the analysis so it seemed quite appropriate that they be used in this analysis as well. Presumably, any intra-task relations found existing in the last four training trials of C-8 and C-13, in particular, would be indicative of the kinds of relationships to be expected in highly skilled, relatively stable performance.

Using, then, the same four-trial sets as had previously been used in computing end-of-training probabilities, the correlations in performance between each pair of parameters within each phase and for each pilot of C-8 and C-13 were obtained. Kendall's τ seemed the most appropriate statistic for this purpose, because in the three-variate (parameter) case a partial correlation can be computed, thus, excluding from the relationship of any two in question any effects of their mutual relationship with the third factor. The outcome of this analysis of the seven phases entailing two or more parameters is presented in table XXIII.

The entries in table XXIII afford interesting possibilities for understanding the approaches which the pilots took in performing the several phases. Presumably, the pattern of the intercorrelations for any particular pilot and phase shows how he organized the task elements, which he was willing to trade off performance in for which, and so forth. An excellent example (almost idyllic) is the performance of the braking and hover sequence (i.e., the lunar landing) by P-133. In one way of describing it, this pilot may be said to have performed the phase as two independent tasks—the rate (in displacement and impact) control task and the displacement-fuel task—trading off performance in the elements of each of these. In contrast, in the same phase P-132 appears to have traded off fuel consumption with displacement rate, while holding these and the other aspects rather independent otherwise. It might be very profitable for understanding complex task performance to obtain detailed information from the pilots as to how they approached the phases and then to attempt to relate that information to these intercorrelations. However, although such an analysis might be very valuable, it is not necessary to the immediate need.

TABLE XXIII

Correlations of Parameters within Phases on the Last Four Training Trials (Kendall's τ)†

Phase/ Comparison		Pilot 81	82	83	131	132	133	All (mean)
TLI	V x P	.000	-.667	-.183	.183	.183	.000	.081
THN	DR x XR	.333 (.250)	.333 (.333)	-.667 (-.707)	-.333 (.000)*	-.333 (.000)*	-.333 (.490)	-.167 (.061)
	DR x F	-.333 (-.250)	.000 (.000)	.333 (.447)	1.000 (1.000)*	.333 (.000)*	.913 (.926)	.374 (.354)
	XR x F	-.333 (-.250)	.000 (.000)	.000 (.316)	-.333 (.000)*	-1.000 (-1.000)*	-.548 (-.634)	-.369 (-.261)
LOI	V x P	.548	-.183	-.183	.183	-.707	-.548	-.148
SDO	V x Y	-.333 (.297)	.667 (.633)	.548 (.657)	.183 (.132)	.333 (.000)*	.913 (.919)	.385 (.341)
	V x P	.730 (.722)	.667 (.633)	.333 (.527)	.333 (.310)	.333 (.000)*	-.183 (.315)	.369 (.418)
	Y x P	-.183 (.093)	.333 (-.202)	-.183 (-.463)	.183 (.132)	1.000 (1.000)*	-.333 (-.414)	.136 (.024)
BH	D x DR	-.333	.333	.333	.000	.183	.000	.086
	D x XR	.183	.000	-1.000	.667	.548	.000	.066
	D x F	.000	-.333	.183	.333	-.183	-1.000	-.167
	DR x XR	-.913	.000	-.333	-.333	-.333	-1.000	-.485
	DR x F	-.667	-1.000	-.548	-.667	-1.000	.000	-.647
	XR x F	.548	.000	-.183	.000	.333	.000	.116
Dok	D x DR	.333 (.527)	.333 (.447)	-.667 (-.334)	-.333 (-.333)	.333 (.447)	.333 (.527)	.055 (.214)
	D x XR	-.183 (-.463)	.000 (.316)	-.707 (-.447)	.000 (.000)	.000 (.316)	-.548 (-.657)	-.240 (-.156)
	DR x XR	.548 (.657)	-.667 (-.707)	.707 (.447)	.000 (.000)	-.667 (-.707)	.183 (.463)	.017 (.026)
EE	D x A	.000 (.316)	.333 (.447)	.333 (.250)	.333 (.333)	.333 (.000)*	.333 (.250)	.278 (.266)
	D x CR	-.333 (.447)	.000 (-.316)	-.333 (-.250)	.000 (.000)	1.000 (1.000)*	.333 (.250)	.111 (.189)
	A x CR	.667 (.707)	.667 (.707)	-.333 (-.250)	.000 (.000)	.333 (.000)*	.333 (.250)	.278 (.236)

† values for partial τ are given in parentheses

* value for partial τ is the logical consequence of the relationship observed, although not obtainable numerically by the usual rules of algebra (numerator or denominator or both equalling zero)

For, what is of primary concern for the methodological issue here under discussion is simply whether and in what degree performances in measured task elements within phases were related. Now, taken all together the correlations in table XXIII do not generally represent a high degree of relationship. There are, of course, instances in which r does reach .667 and greater but these are relatively infrequent compared to the instances in which it is .333 or less. Furthermore, and more crucial still, is that a clear organization of task elements, as indicated by some strong dependencies (high r) and some independencies (low r), seldom emerges. What is much more frequent is a set of mutually low correlations representing rather weak relationships among the task elements. Rather evidently, one can hardly hope to base a useful correction of $\bar{p}$ and $\dot{p}$ computations on such weak contingencies as are oftentimes indicated. Thus, any hopes for developing more precise expressions of the phase performance indicants must be restricted to only those individuals and phases in which moderately strong relations are in evidence. (There are 20 cases in the total of 40 in which the greatest correlation is .657, or more.) But even correction in many of these would not be especially useful because of the considerable independence still involved in the relationship of the most strongly correlated pair of parameters.

It is apparent also that even less hope exists for finding a suitable expression for each phase which could be used for all pilots and so, would offer great computational convenience. Casual cross-comparison of row entries as well as the magnitudes of the row means in table XXIII shows that without question the pilots were not consistent in their approaches to a given phase. The largest mean is -.647, indicating an inverse relation between fuel and displacement rate in braking and hover, and even in this case there is an exception (P-133). Even among the 20 cases yielding moderately large maximal correlations there is inconsistency in the patterns to be observed.

Consequently, it was concluded that although expressions correcting the computation of $\bar{p}$ and $\dot{p}$ for certain phases and certain pilots might be feasible and bring about improved accuracy—the infrequency with which significant improvement could be gained and the complications involved do not justify doing so. Instead, it seemed wiser to merely recognize, as has been done, that $\bar{p}$ and $\dot{p}$ for certain phases and individuals are likely to be somewhat biased by the failure to take account of the contingencies among task elements. The exact character of the bias can only be discerned from examination of the pattern of correlations and is unique to each instance. In only a few such cases is the resulting error likely to be of any size numerically. In no instance is it considered sufficient to influence comparisons of a given pilot's performance at various times because the bias in compared values presumably would be consistent. These statements have reference, of course, only to C-8 and C-13. Because C-4 and C-9 apparently did not near their peak skill, it is unlikely their performance reflected a stable organization and that biases due to task element dependencies exist in it.

In summary, then, certain analytic procedures used along with certain data collection procedures prompt serious concern in the interpretation of the data. There are others which might be supposed would cause a problem but which on examination are considered to give little difficulty. Among those of concern, the likely inaccuracies in estimation of p resulting from the use of such a small sample, particularly the influence of occasional extreme values, necessitates watchfulness in the interpretation of deviant p values. It is especially

likely that the original capabilities of P-132 and P-133 are somewhat underestimated because of extreme values in the reference last four trials of training and that these extreme values influenced the estimates about their test mission and retraining capabilities as well. Furthermore, in having been obtained from the normal probability function rather than Student's t distribution, all probabilities of meeting criteria are generous whereas probabilities of meeting former $p_{.95}$ levels are underestimated. Test mission probabilities also tend to be overestimated by the use of a "borrowed" variance estimate in calculating them. Procedural details of the test mission and in retraining may have operated to degrade performance in them somewhat. The disruption in plans almost certainly degraded P-81's retraining performance. On these grounds it may be supposed that with some individual deviation the capabilities of C-8 and C-13 to meet criteria are generally overestimated in the results presented, while their capabilities in test and retraining to meet former $p_{.95}$ levels are generally underestimated. However, in considering $\dot{p}$ and the more inclusive indicants involving it additional account must be taken of the disproportionalities inherent in $\dot{p}$, as compared with $\bar{p}$, and the occasional modest inaccuracy in estimates resulting from treating task elements as independent.

Although some sampling bias is considered likely in retraining, owing to the selection of the best four trials for analysis, this effect is discounted as roughly compensated for by diminished motivation. The use of dual criteria for retraining performance complicates interpretation but is not a source of bias. In general, the data do not suggest that any bias of consequence is likely to have resulted from asymmetry in distribution of cases—that is, failure to reflect normality in this way. Missing data is not a problem.

The Problem of Generalizability

Having considered the more specific data collection and analytic procedures and their probable impact on the results, it remains to deal similarly with certain more general issues affecting the broader implications of the study. These are matters which strongly influence the range of applicability, or generalizability of the findings. Among them are such varied concerns as statistical significance, intervening activities, skill differentials, the use of simulation, and test personnel motivation. Each of these will be treated briefly, in turn, in the following discussion, the primary intent being to express the position to be taken in the subsequent statement of findings. It is not intended that the comment on these topics included here be a comprehensive treatment—for that is beyond the scope of this report. Furthermore, it should be recognized at the outset that alternative views to those presented are possible because these are debatable matters.

Statistical Significance. In the preceding brief description of analytic methodology it has already been noted that conventional tests of significance for skill retention effects are simply not feasible in view of the structure of the data. It was not possible to combine the performance of even two pilots to arrive at the necessary error estimate for such a test because (in the language of experimental design) all individuals represent unique treatments. Furthermore, although a specialized adaptation of the analysis of variance technique might be applied, the complexity of the data structure makes doing so inadvisable. Therefore, it seemed best to merely examine the pattern of results obtained with an eye to the internal coherence of the data and correspondence with reasonable expectations concerning the likely effects of known factors involved.

Now, of course, there are those who take the position that such a course of action is not acceptable—that not to determine levels of statistical significance for differences observed is tantamount to unscientific. After all, if one does not know the level of significance attaching to an observed difference how can the meaning of that difference be known? There is no fundamental quarrel with this position—so far as it goes. Most certainly, a real advantage in precise interpretation and enhanced meaningfulness does accrue from the expression of statistical significance. However, the difficulty with this simple view of data interpretation is that it does not go far enough to comprehend the realities of observational complexity in many situations of interest. Also, it does not allow for the investment of meaningfulness by other methods, and presumes that only the statement of statistical significance can serve that purpose. And yet, oftentimes it is quite important to evaluate in some way the regularities of complex data. Should all hope of doing so be given up for the lack of a ready model for determining the likelihood that a particular effect occurred by chance?

In a broader view of the data interpretation problem the most appropriate answer must surely be, No! Although *t*-tests, F-ratios and the like have great utility when they can be applied, to argue that without them nothing can be done is, frankly, to denounce a great variety of data as of no use. It is to sacrifice a host of potentially useful observations on the altar of significance in an unthinking fervor of statistical reverence. Instead, it seems that, with all due regard to the difficulties involved, the effort can and should be made in any case to examine the organization of the results obtained. For, a degree of meaning also may be attached when a set of interrelated observations also shows a high degree of internal coherence or an expected pattern of magnitudes. What, after all does or should an investigator do on encountering a collection of *t* or F values not quite significant, but all readily interpretable on extra-statistical considerations? Should he conclude that no meaning is to be derived from the results? To do so seems not only unwise, but wasteful.

Accordingly, the position adopted in this report is that if a collection of observations follow rather consistently reasonable expectations concerning their order of magnitude, it should be concluded that the existence of the effect is thereby demonstrated. However, because the degree of consistency required for acceptance of a particular conclusion is a matter of judgment, a sufficiency of information is always presented to enable the reader to himself examine the consistency involved and, thus, to evaluate the merits of the conclusion given.

On the other hand, the adoption of this policy should not be taken to imply that tests of significance on these performance data are believed to be impossible. Even though the data are presented in an unusual probabilistic fashion it is still rather obviously possible to apply counting or frequency methods to them to arrive at some rough indication of significance and, in turn, broad generalities. For example, one can readily calculate the probability of, say, six out of six crewmembers not reaching the same p_c in test as in training. Such calculations were not considered necessary for interpretation, but some readers may wish to do so for their own interest. Finally, beyond these offhand methods, more refined techniques for testing the significance of differences in the same individual's performance—say, between training, test, and retraining—also are possible. But, because further development and study of them before actual application is considered desirable, no attempt was made to use them in this analysis.

Intervening Activities. Turning attention now to broad issues concerned with data collection procedures, a so far undiscussed concern of great possible significance is that of the activities intervening between training and skill retention test. In early consideration of memory and forgetting psychologists inclined to the supposition that the mere passage of time was sufficient to account for the degradation in skill or recall noted subsequently. However, recent studies in which the nature of the activities engaged in during the retention period have been varied have cast serious doubt upon this position. Thus, today it is commonly held that these intervening activities have much more influence on subsequent performance than the length of time per se.

Mindful of the importance, then, of intervening activities for a close interpretation of the data, upon reporting for the skill retention test each pilot was asked informally to describe in a general way what he had been doing since the completion of training. No attempt was made to estimate specific numbers of hours involved or to go into sensitive details. Each was asked also to what extent he thought these activities were helpful in maintaining his capability to perform the lunar landing mission.

In brief, the consensus of their responses was that each had engaged throughout the period in normal test pilot activities, including conferences, analysis of engineering details, test planning, and flight in high performance jet aircraft. Furthermore, they felt that, in particular, their periodic flights in high performance aircraft brought substantial benefit in maintaining their capability to deal with a complexity of factors under time pressure, as was required in certain portions of the test mission. They did not feel that this benefit accrued from specific similarities in task requirements but rather from the maintenance of more general habits as might be characterized as alertness to requirements for time-sharing and rapid shifts of attention, and for rapid response. Several, when later asked specifically, did not feel that any particular dissimilarities in required tasks had evident negative effects on their performance (they are used to dealing with such dissimilarities) and confirmed their earlier convictions.

If the pilots' views of the effects of the intervening activities are taken at face value (and there is no evident reason they should not be) then what implications do they have for the study findings? Surely any statement regarding the degree of skill loss encountered or to be expected must be qualified by reference to the presumed benefits of these intervening tasks. This implies that, other things being the same, pilots whose intervening duties do not involve tasks offering similar beneficial effects can be expected to perform less well. It also follows that the present estimate of skill retention is a generous one compared with what might be derived from other circumstances involving different intervening activities. On the other hand, that is not to state that even greater skill retention could not be obtained with some still more beneficial intervening task requirements.

Skill Differentials. Still another general feature of study procedure having great potential bearing on the interpretation of the results is the level of skill attained prior to test. This factor has, of course, already been considered in the presentation of training performance and the necessity of distinguishing C-4 and C-9 performance from C-8 and C-13 performance in this respect seems established. Furthermore, it now appears that even the best trained crews, in general, had not quite reached the limit of their capability. However, beyond

the gross differences in crew capabilities in the two study phases there are also the more subtle differences in capabilities of crews and individuals within study phases. The pertinent concern here is what do these differences imply, particularly in the case of O-8 and O-13 of primary interest, for the conclusions to be drawn.

In brief, it is contended that the differences in individual pilot capability to perform the various mission tasks as well as the differences among pilots and crews in capability to perform them greatly reduced (or eliminated) the possibility of showing the exact effects of both retention interval and testing order. The pilots of O-8 and O-13 had indeed reached similar levels of skill as shown by the overall indicants but they were still not the same and they differed considerably on the tasks of relative expertise and deficiency. As a result the level of original skill—known to be a powerful determinant of retention—must be assumed to have operated variously to either enhance or degrade the relative performance of the individuals and crews. Since level of learning effects are relatively great it may be supposed that they overshadowed such lesser effects as might otherwise have been observed. Thus, for example, it may well be that differences in the losses of P-81 and P-131 on test might have been noted had these pilots not differed in a fashion parallel to the hypothetical effects of retention interval. If P-131 had been exactly as skilled as P-81 he might have experienced even greater loss than he did.[4]

That it was possible, even though skill differentials did exist, to distinguish the apparent effect of testing order such that the first-tested did less well, is a testimony to the relatively strong beneficial effects of opportunity to observe and participate in a prior performance. But, that retention interval had no clearly distinguishable effects implies only that it did not have as strong an influence as order of test or original learning level—not that it had no effect. Thus, it must be understood that because of the lack of exact control over skill level no conclusion concerning retention interval beyond its relative strength as a variable can be asserted.

Simulation Versus Operations. One final issue relating broadly to the fundamental nature of the study is the question of simulation versus real operations. This matter is popularly considered of such significance that for many readers the viewpoint taken on it will primarily determine the credibility of anything else which may be asserted about the study. And, of course, general opinions on the appropriateness of using simulation techniques in aerospace system research and design range all the way from near rejection and grudging willingness to see them used in very preliminary development work to enthusiastic acceptance of them as of great value throughout all phases of research and development, including the final determination of operational readiness.

---

4 Somewhat confusingly, and at variance with conventional expectations, at high reliability levels a greater degree of skill may be associated with greater (not lesser) losses in reliability. The man who is most reliable has, in a sense, most to lose.

This study is, frankly, prefaced upon the belief that simulation techniques afford a useful means of gaining data on and indications of human performance capability otherwise not obtainable or obtainable only at great cost and effort. It is not the intent of these remarks to attempt convincing those who lack this degree of acceptance of simulation methodology that they should adopt a more positive view. But, in order to be in the best possible position to realistically interpret the results obtained it is necessary to consider the probable consequences of the simulation methodology employed. For, to accept results obtained by means of simulation without regard to the limitations inherent in the method would constitute an indefensible, blind acceptance. In fact, it is only when the results obtained by simulation are interpreted with care to insure that account is taken of the influences of the method on the data that the utility of the method can be realized.

What, then, are the probable effects on the data deriving from their being taken from a simulated performance rather than a real operation? Or, stated another way, how would pilot performance have differed if the test mission had been a real lunar landing mission and the retraining performances had been imbedded in real lunar missions? For that matter, it may be questioned, as well, how the estimate of end-of-training capability would have changed if real missions had been used instead.

In formulating and considering answers to these questions it is convenient to distinguish three main kinds of ways in which performance could be influenced. In the first place, specific task features may differ between the simulated and real situations. A simulated situation designed to replicate as exactly as possible the real situation (as is the traditional approach in the design of training simulators) might be expected to produce the same results as the real situation. However, it is the fact that simulators never quite yet reproduce the real environment: typically there are still missing from them complexities of the visual world and of the motion dynamics, if not some of the simpler operational aspects. Consequently, it may be supposed that as a result of such specific dissimilarities performance will be modified—either enhanced or degraded, depending on the nature of the differences.

In this study, although the simulation is considered to have been well done and as nearly complete in task specifics as current technology permits, it nevertheless departed to some degree from reality, particularly with respect to motion dynamics and visual cues in the lunar landing. Furthermore, many of the environmental factors of space flight, such as weightlessness and radiation, were not included and realistic attention to life support tasks was not required. Finally, perhaps most important of all, while the specific task details were chosen to duplicate in a general fashion early planning for the real Apollo even these, now somewhat obsolete task details, were modified deliberately in the direction of greater manual control.

The net effect of all these specific differences is, of course, a matter of conjecture but some reasonable implications can be drawn. The most significant is probably that the performance observed cannot be taken as exactly indicative of that to be obtained in any particular real system, including the Apollo. Rather, to the extent otherwise valid the obtained reliabilities may be taken only as broadly indicative of the orders of magnitude which may be expected in a mission of the nature used. No precise generalization to a particular system may be based on these data unless it is established on close analysis that the details of the tasks of interest were exactly comparable.

Furthermore, although the absence of environmental stressors and life support requirements reasonably may be expected not to have biased the performance from that of normal operations, there is always the possibility in real operations of a problem in these areas. Such an occasional problem could result in a degradation in other aspects of performance. That happenstances of this nature were missing from this simulation, suggests that the obtained performance estimates are somewhat generous, but in an unpredictable fashion as to what portions of the mission might be involved. On the other hand, the specific deficiencies in simulation detail, particularly for C-4 and C-9, probably operated to make the braking and hover phase, and perhaps docking, unrealistically difficult—certainly this is the view expressed by the pilots.

Then, in addition to the specific differences existing between simulated and operational situations, there are also nonspecific differences arising from the performance contexts. Two kinds of contextual factors may be usefully identified and separately considered.

The first of these is the possible effect of disturbances in performance continuity which often occur in a simulated operation. Unlike the real mission which unfolds in continuous fashion, each particular phase following in an orderly way upon the previous, the simulated mission is often characterized by unrealistically abrupt changes in task requirements, by the omission of less-important or nonessential tasks, by the elimination of longer time periods in which specified mission activities are not required, and even sometimes a juxtaposition of task requirements. All of these typical departures from realism were involved in this study and so it is essential to inquire how they may have affected the results.

In general, with respect to the test mission, it is supposed that the net affects of these deviations from realism were to somewhat degrade performance. All of them may be described in psychological terms as disturbances which tend to break or disturb the development of the pilot's set for performing a given task by prompting his precipitous attention to the new activity without the benefit of the usual preparation time and prior cueing events which he may have learned to depend upon. In contrast, training and retraining performance would perhaps be much less disturbed in this way—at least that following the initial performance of a given phase. Indeed, the opportunity to repeat a performance immediately may have resulted in an enhancement due to the maintenance of set and better immediate recall of task details (especially in more difficult phases) up to the point of fatigue and boredom. That some boredom probably occurred in the last portion of retraining has already been mentioned.

The second kind of nonspecific difference resulting from context concerns the emotional and motivational concomitants of performance. It is just plainly and irrevocably true that simulated operations (including this one) generally do not involve the pilot in the same degree of risk to personal safety and potential loss from failure as do operational missions. Furthermore, on the realization of their simulated nature, exactly the same degree of importance is seldom attached to success in simulated missions, and even if it is the reasons are different. Consequently, in most cases the emotional responses as well as the motivations of the individual performing a simulated mission are different from those of an individual performing a real mission.

Now, on the strength of these differences in emotions and motivation it is often supposed, therefore, that the individual will perform the simulated mission somewhat less adequately than the real mission. This follows from the common supposition that greater personal concern results in superior performance. However, the difficulty with this simple view is that considerable experimental evidence shows that there is not a direct relationship between performance adequacy and personal concern. Instead, it appears that increasing personal concern brings performance improvement at a diminishing rate up to a certain point beyond which it brings performance degradation at an increasing rate. There seems little doubt that increasing motivation and emotional responses to a task beyond a certain point is not only of no value, but may be detrimental. But, within a considerable range about the point of optimum, changes in personal involvement result in only slight changes in performance. Thus, in order to estimate the probable relative consequences for performance of motivation and emotionality in the simulated mission, contrasted with a real mission, it is necessary to judge where, with respect to this optimum range, the respective degrees of such involvement are placed—not simply that they differ.

Previous experience with space operations suggests that the personnel become keenly and personally involved and experience some anxiety, but that these factors do not reach such proportions as to much degrade performance. This suggests that space missions are typically performed within the optimum range of personal involvement, as is to be expected because considerable development effort and operational planning is directed toward this goal. What, then, can be said of the involvement of the test personnel in the present simulation study?

Again, as a matter of judgment and estimation, it seems likely that the pilots' motivation for excellent performance and their anxiety in the retention test mission were such as to place them within the optimum range. Because of the inherent challenge to their professional skills the test afforded and the considerable attention given it on the part of certain professional associates the pilots were quite eager to perform well. They viewed the test as having at least some implications for their career. Accordingly, so far as motivational and emotional concomitants are concerned their performance in the skill retention test mission is considered comparable to that of an operational mission.

On the other hand, the same cannot be asserted about their retraining and, perhaps, about their end-of-training performances. While, they no doubt began retraining with considerable interest in achieving their best performance the situation was not the same—they had, in a sense, already taken and passed the test and the test was anticlimax. Hence, it is supposed that while their personal involvement probably continued at nearly the same representative effectiveness level for the first block or two of trials on the first day of retraining, thereafter it diminished to the point of noticeable effects. On similar grounds it seems doubtful whether the pilots performed the very last portion of training (having long before demonstrated regular capability to meet the prescribed nominals) with the optimum and representative degree of involvement. Thus, it is concluded that although motivational and emotional factors attributable to the simulated performance situation probably did not bias the skill retention test they probably did result in some underestimation of capability at the end of training and after extensive retraining.

In summary, with respect to the problem of generalizability, the position adopted for the purpose of formulating conclusions is simply this: The lack of traditional tests of significance does not preclude the drawing of useful conclusions providing this is done with care and the reader can evaluate the basis for the conclusions. The activities of the pilots intervening between training and test had generalized beneficial effects on their test performance which could, but might not be, duplicated in other skill retention circumstances. The lack of exact control of level of original skill rendered the study incapable of showing the precise effects of length of retention interval but did not overshadow the gross effects of retention per se and individual testing order. And finally, the use of a simulated situation probably tended to reduce the level of performance in training and in retraining, as compared with a real mission, but did not much affect test mission performance.

The Considered Findings

The foregoing discussion of factors which should be considered in interpreting the results obtained is not necessarily exhaustive. Nevertheless, it does make evident that there were a number of factors operating which in some degree did or may have influenced either the performance itself or the analytic results and, therefore, led to either over- or underestimation of pilot capability on occasion. Some of the factors identified are of such a nature as to bring about fluctuations by parameter and phase in either direction in a way difficult to guess upon in any particular instance. But, since no clear basis was found for supposing them to result in a definite bias one way or the other it is reasonable to discount them from further concern in an overall view of the results. They may well have affected certain p, $\bar{p}$, and $\dot{p}$ values, but they probably had little affect upon $\bar{\bar{p}}$, $\bar{\dot{p}}$, $\dot{\bar{p}}$, etc.

On the other hand, there were identified a few other factors which probably did introduce a bias (though not necessarily a consistent one) rather than merely fluctuations in p values. These, naturally, influenced the overall indicants in a parallel fashion as well and, therefore, they must be considered in interpreting the results. Thus, omission of certain requirements (unpracticed emergencies, management of life support equipment, etc.) tended to result in performance overestimation throughout the study. In addition, the use of the normal, rather than the t distribution, also brought an overestimation of capability to meet criteria and an underestimation of capability to meet former $p_{.95}$ performance levels. The relative lack of precise goals and precise feedback in the procedures used with C-4 and C-9 consistently placed these crews at a disadvantage compared with C-8 and C-13.

Still other biases were identified which influenced the results for only one or two but not all three main portions (training, test, and retraining) of the study. In training the lack of motivational and emotional concomitants of a real mission operated to generally depreciate estimated capability. In the test mission the break in set (resulting from time compression) and perhaps some other procedures tended to depreciate the attained reliabilities, whereas the use of variance estimates based on training data tended to elevate them. In the retraining, again the lack of realistic motivation and emotional responses and perhaps certain procedures probably degraded performance, but these effects are at least partially compensated for by making the calculations with reference to the best four trials.

The results for certain pilots, in particular, also probably are biased. Because of the extreme values recorded at the end of training, it is quite likely the estimated capabilities of P-132 and P-133 to meet criteria in training and in the test mission are spuriously low and their actual gains (if any) in retraining over training less than indicated. The retraining performance of P-81 was almost certainly adversely affected by the other requirements imposed on him.

Evidently, some of the biases operated to underestimate capability while others operated to overestimate it and so, to some extent at least, they tended to cancel each other out. To what extent they did cancel out depends upon the particular combination of amounts and directions of bias in particular portions of the study and in the two types of analysis (criterion and $p_{.95}$ level). In general, after careful weighting of positive and negative influences according to their estimated magnitudes, it was concluded that the overall estimates given (i.e., $\bar{\bar{p}}$, $\bar{p}$, $\dot{p}$) of capability of C-8 and C-13 to meet criteria in original training and in the test mission are somewhat overestimated; while capability to meet criteria in retraining is slightly underestimated. (The values for P-132 and P-133 in training and test and for P-81 in retraining were discounted from this generality as spurious.) These net effects are judged to be quite small relative to the reliabilities reported and not to exceed about .01 in $\bar{\bar{p}}$. In particular, the probabilities reported for training performance are considered to be very close to those which might be obtained in a real mission. Presumably, the values for C-4 and C-9 in training, test, and retraining are similarly affected. However, in having had less effective task goals and performance feedback they probably performed less well than they might have throughout the study. This is a bias only insofar as their performance is compared with that of C-8 and C-13.

In a similar fashion it was concluded that the overall probabilities of C-8 and C-13 pilots realizing their individual $p_{.95}$ levels in the several aspects of flight control in the test mission and in retraining are underestimated. In general, the amount of underestimation or net bias in $\bar{\bar{p}}$ in these instances was judged to approximate .03. The parallel overall probabilities of C-4 and C-9 pilots may have been similarly affected, although possibly to a lesser extent because of their lesser original skill.

On the basis of this evaluation of the influence of the various interpretive factors it seems unlikely that any inaccuracies in the results are of sufficient magnitude to invalidate the main comparisons of interest. For example, if the test capability of C-8 and C-13 is overestimated more than original capability then the reality of the observed decrement is more, not less, assured. Similarly, the pattern of supposed inaccuracies suggests that by additional performance in a real situation these crews would have recovered their original capabilities to meet criteria even more quickly than they seemed to in this study. Furthermore, all of the inaccuracies described are sufficiently small as to permit a rather precise impression of the actual reliabilities to be expected in a real mission. It is only necessary to make the indicated minor adjustments in reported values. Therefore, it is contended that the main findings, as presented in the section on results, should be taken at face value so far as differences among training, test, and retraining performances are concerned, with minor qualifications added about probabilities to be expected in a real mission.

Accordingly, the considered findings are as follows:

1. Original Capability:

a. By the end of five weeks training the pilots and crews achieved varying levels of reliability in performing the nine mission phases according to hypothetical system criteria.

b. The first two crews tested (C-4 and C-9) were generally much less able to meet the criteria than the second two crews. Highly reliable (.999) performance was seldom demonstrated by them on any portion of the mission and even the best performer among them showed a likelihood of less than .850 of performing within the criterion on any parameter ($\bar{p}$). The likelihood of meeting all criteria (i.e., $\dot{p}$ or mission success) was never more than .002. Because these crews evidently were not fully trained, their data are discounted in the present study as incapable of supporting useful generalizations about skill retention in space missions.

c. The second two crews (C-8 and C-13) attained quite high indicated reliability levels. Estimated likelihoods of meeting the criterion on any parameter ($\bar{p}$) ranged from .932 up to .990 and of meeting criteria in all parameters ($\dot{p}$) from .294 to .817. However, the lower two sets of values (for P-132 and P-133) are considered underestimates due to the inclusion of unrepresentative scores, whereas all other values are considered overestimates of true capability in a real mission of this nature. The extent of this overestimation is judged to be in the order of .005 in likelihood of meeting any particular criterion and of .10 in likelihood of meeting all criteria. Even these apparently well-trained pilots differed somewhat in capability in the several measured aspects of flight control, the different phases, and the overall mission.

2. Test Mission Capability:

a. In the test mission, performed approximately 8 to 13 weeks after training, the well-trained crews (C-8 and C-13) showed an estimated capability to meet the criteria somewhat lower than their original capability to meet them. The first-tested pilots evidenced considerable loss, but some of those tested second and third showed a slight, probably spurious, gain—the crew averages, consequently indicating mild losses. As was expected, the same pilots showed even less likelihood of reaching former highly reliable ($p_{.95}$) levels of performance in individual parameters. Evidently, lack of direct practice over a period of 8 weeks or more resulted in a definite loss in skill which was to a considerable extent recovered by the opportunity to observe or participate once or twice in another pilot's performance of the mission. However, a difference in retention over 8 and 13 weeks was not demonstrated—perhaps because of differences in original learning—which indicates that such variations in retention intervals that long may be of little significance. The levels of reliability derived from the first test performance of around .92 of meeting the criterion in any parameter and of around .02 of meeting the criteria in all parameters are considered to slightly overestimate real mission performances under like circumstances of previous training, intervening activities, etc.

b. The incompletely trained crews (C-4 and C-9), tested after 4 or 9 weeks, showed generally parallel effects of the lack of practice on capability to meet criteria—but to a lesser extent and somewhat inconsistently. This differ-

ence is attributed directly to their smaller amount of original skill. However, they showed about the same loss in capability to perform at their own former levels of high reliability as the well-trained crews. This suggests the interesting possibility that losses in relative reliability (i.e., probability of realising former levels of highly reliable performance) may be little influenced by original skill attainments.

3. Capability with Additional Training:

a. In the course of 3 days of additional training the well-trained pilots of C-8 and C-13 again showed slightly varying capabilities in the several measured aspects of flight control, the different phases, and the overall mission. They tended to be less able to meet the criteria for separation and deorbit, brake and hover, and earth entry phases than they were able to meet the criteria for transposition, lunar orbit and docking. Even in the very first four trials of retraining four of these pilots approximated their original capabilities for meeting the criteria and in their best four-trial blocks in the various phases all six exceeded those capabilities. Thus, original capability to meet system criteria was quickly regained. That the original capability to meet criteria was improved upon implies (especially because retraining values are believed to be underestimated) that these pilots had not originally reached the limits of their potential. However, even at their best, these pilots did not demonstrate in retraining their full original capability to perform all the measured aspects of flight control. Their efforts to maximize performance with respect to criteria with accompanying frequent shifts between training and retraining in relative performance on aspects within phases, along with the sensitivity of this analysis to such shifts, is considered the probable explanation. Such changes in strategy, which are costly of overall reliability, are not reflected in the criterion-reference analysis because they are generally beyond the arbitrary maximum discriminable reliability of .999.

b. In the course of two to eight additional training trials on a single day, the less well-trained pilots of C-4 and C-9 were able to demonstrate capabilities for meeting the system criteria comparable to or exceeding their original capabilities for meeting them. However, like the pilots of C-8 and C-13, they did not demonstrate their full original capability to perform all the measured aspects of flight control, although they may have come somewhat closer to doing so. They continued in retraining to be distinctively less able to meet the criteria, even though they performed at least as well relative to their own former levels. Thus, although operating at lower levels of capability with respect to criteria, the pilots of C-4 and C-9 responded to additional training in about the same way as did the pilots of C-8 and C-13.

IMPLICATIONS

The detailed outcomes of the study just given are, of course, valuable in their own right. But, of even greater interest, perhaps, are the implications they carry specifically for the design of space systems, and more generally for the measurement and prediction of human performance. They are also suggestive of directions in which additional research on skill retention and on related measurement methodology should go. These apparent implications will be briefly presented, in turn, in this concluding section.

Space System Design

Skill Retention in Long Missions. As described at the beginning of this report, the primary intent of the study was to evaluate the degree to which skill in critical tasks might be retained in space missions of extended duration. For, in designing such systems one would certainly like to know whether forgetting on the part of the operators would be enough to degrade the probability of mission success below acceptable levels. If forgetting in that amount were predicted then an obvious design strategy would be to seek to reduce or avoid the degradation, possibly by the modification of the mission profile or the mission tasks or by the use of special training procedures. But, it is first necessary to know whether skill retention is likely to be a problem.

The findings of the present study indicate that a carefully qualified or guarded position on the skill retention question is in order. If the mission test performances of P-81 and P-131 (the critical test personnel) are taken at face value then it must be concluded that skill retention difficulties can indeed occur. P-81, tested after 8 weeks, was able to perform most of the tasks required within the nominal criteria, but he did not meet them in two parameters. P-131, tested after 13 weeks, also was able to perform most tasks within criteria, but he deviated greatly from them in three of four critical parameters in the lunar landing. (On a real mission he would have sheared the landing gear, at the very least.) Because of these difficulties both pilots' estimated reliabilities in meeting criteria dropped noticeably from their former levels to reliabilities which would be unacceptable for a real mission. (P-81 dropped 6% to .931 in $\bar{p}$, 14.8% to .827 in $\bar{p}$ and 93.5% to .048 in $\dot{p}$ while P-131 dropped 7.8% to .899 in $\bar{p}$, 9.3% to .858 in $\bar{p}$ and 100% to <.001 in $\dot{p}$.) The estimated probabilities of mission success (i.e., of meeting all criteria) by not exceeding five chances in a thousand are particularly discouraging. Quite evidently the requirement in a space mission to remember a critical flight control task over 8 weeks or more poses a problem worthy of special design attention.

However, in addition to demonstrating the reality of the skill retention problem, the present data provide some insights on the severity of the problem and indications as to how it might be handled. Of particular interest in this connection are the test mission performances of the pilots tested second and third and the retraining performance of P-131. (The retraining data from P-81 are discounted for reasons given earlier.) Although pilots 82, 83, 132 and 133 had at the end of training attained approximately the same level of skill in meeting the criteria as P-81 and P-131, all but P-82 showed a slight gain in test over training and P-82's loss was relatively small. Consequently, there seems little doubt that immediately prior observation or aiding of another once

or twice in performing the critical tasks can bring about a considerable improvement in performance after 8 or more weeks without direct practice. Furthermore, in just four practice trials following the test mission, P-131 equaled or exceeded his own former capability to meet the nominal criteria (at least to a probability of .999) in six representative phases of the mission. On the assumption that the chosen criteria are reasonable ones for such systems, both of these findings imply that the skill retention problem need not be severe. They suggest that the degradation in mission reliability associated with recall over at least 13 weeks can largely be avoided in instances where intervening activities serve to maintain an alertness to requirements for time-sharing and rapid response and when opportunities exist to practice or even rehearse and observe the critical activities before they are performed in the mission. Apparently, the preexposure to the tasks need not be extensive, but it must be more than merely an examination of typical handbooks and checklists and verbal discussion.

The comments and discussion of the debriefings, conducted informally after the test mission, lend further support to this view. In essence, the pilots' comments then (and earlier) emphasized the desirability from their point of view of an opportunity to enter the crew station beforehand and refamiliarize themselves more directly with the displays and controls. They felt this would have alleviated considerable misgiving on their part about performing the tasks adequately. Beyond this, in response to direct questioning, the pilots indicated a belief that the prior opportunity to observe or rehearse the more difficult phases would be very helpful and the second- and third-tested pilots confirmed this from their own experience. Most agreed that the casually-suggested possibility of a preview by means of an over-the-shoulder filming of action of the more complicated phases (perhaps narrated themselves for their own use) might be quite helpful. However, P-81 took exception to this because he felt that only the opportunity to actually operate the controls would be helpful to him.

Of particular interest in the debriefings were the comments of P-131, about whose difficulty discussion naturally centered. He stated that he had proceeded with the braking and hover sequence very cautiously in an effort to do well and did not realize until too late to fully recover, the implications of his excessively slow rate of descent. He described this as a problem not encountered in training, thus seeming to imply training inadequacy, and said that now being aware of such a possibility he would not likely repeat his mistake. However, a somewhat different, but not contradictory, interpretation is possible, for, it may well be that P-131 did not encounter the problem in training because at that time his greater awareness of the timing factors involved and familiarity with the time-sharing requirements did not lead him to it. Perhaps it was the forgetting of just these task features which occasioned the problem. Certainly that interpretation is in keeping with some of the other pilots' comments on the retention test as well as with laboratory studies on the forgetting of complex motor tasks (Trumbo et al, 1965).

Thus, regarding the question of skill retention in space system design, it seems fair to conclude from the present study that without special design attention operator reliabilities in critical tasks performed after 8 weeks or more without prior direct familiarization are likely to be unacceptable. This does not mean failure to meet criteria necessarily will occur—only that it is unacceptably likely to occur. But, if judicious use is made of training techniques prior to the more difficult critical task performances which afford a refamiliarization with the task (particularly the time-sharing and timing requirements) much of the

expected degradation in reliability can be avoided. However, this should not be taken to mean that the full original capabilities of the operators will be regained by such expedients—in fact, the present data suggest that much more would be required for that. Furthermore, this conclusion rests directly on the assumption that the intervening activities required do not give rise to specific interference effects (negative transfer to the tasks of concern). Accordingly, the possibility of adverse transfer from intervening activities along with the choice of appropriate refamiliarization techniques must be given close attention in design if the problem of skill retention in long missions is to be avoided. In short, skill retention will not be a problem if proper attention is given to skill retention requirements in design and operational planning.

Pilot Reliability. In addition, and incidental to the main concern for skill retention, the study also provides the possibility of insights into pilot reliability beyond the report by Grodsky et al (1966b) of the original NASA-Martin study. This results from two factors: the analytic methodology employed and the evident gains in retraining over the original training performance. In the retraining portion of this skill retention test the pilots were at their very best and the present analysis characterizes that best performance in a way that is quite different from the method used by Grodsky et al. What, then, do the results obtained indicate about pilot reliability?

If it is assumed once again that the hypothetical criteria used are reasonable for space systems—then a fair conclusion on the basis of the retraining values reported for O-8 and O-13 in table XIII should be possible. Taken at face value the indicated reliabilities are reassuringly high, for among the six pilots the minimum values obtained are .991 in $\ddot{\bar{p}}$, .964 in $\bar{\dot{p}}$, and .789 in $\dot{p}$. Obviously the chances of performing a parameter or a phase successfully (i.e., without exceeding criteria) seem quite good and even the likelihood of performing the whole set of phases successfully seems almost as good (the second lowest $\dot{p}$ is .928). The average probability of overall success ($\dot{\bar{p}}$) for the six is .936. Furthermore, the evaluation of possible biases suggested that these values somewhat underestimate reliability in a real situation.

However, there is another factor so far unmentioned which tends to inflate these values, particularly for $\dot{p}$, beyond what might be expected in a full mission. After all, these probabilities for retraining are based upon only six of the mission phases and do not represent a complete mission. Inclusion of the rest of the mission phases would tend to lower mission reliabilities even though parameter, and possibly, phase values could be increased by very good performance. It is an inherent feature of joint probabilities that when other factors are added the product can only equal or become less than the original; it can never exceed the original. Thus, while difficult to estimate, the net result may be that the indicated overall values ($\ddot{\bar{p}}$, $\bar{\dot{p}}$, and $\dot{p}$) are very close to what might be observed in a full mission. Is, then, a probability of around .936 of mission success ($\dot{p}$) high enough for a space mission? Some may doubt that it is high enough and appropriately wonder whether pilot reliabilities in lengthy and involved missions, such as a lunar landing mission, are likely to be disappointingly low and to pose, when taken jointly with equipment reliabilities, a worrisome (or even grave) risk of mission failure. In view of present planning for space systems, to have to answer, Yes! would be chagrining indeed.

Fortunately, a negative answer seems most appropriate for two reasons. First, it is doubtful if the pilots of C-8 and C-13 at their best had reached the full limits of their potential. The still considerable variability evident in their performance as well as informal tests of the training decision rule (referred to earlier) on other data suggest that under a still more stringent training regime (including revised task goals and feedback on performance) all of these pilots could reach new highs in performance excellence. Operational training may be or, at least, can be superior to that given the test personnel in this study.

Second, the so far unquestioned assumption as to the reasonableness of the criteria used may still be doubted with, perhaps, good reason. It is not so much that a question can be raised about one or more of the actual values used: they can be easily remedied by the justification and choice of new values, should anyone care to do so. Rather, the possible problem with the criteria seems to come directly from the very process of using them at all. With respect to many aspects of flight control setting a criterion, other than an extremely broad one, is an arbitrary process. The characteristics of the machine with which the operator interacts most commonly are such that the effects of his error vary continuously in magnitude over quite a range before catastrophic degradation in effectiveness occurs. The overstress point is often well beyond any human performance criterion of interest and so frequently the criteria used can be exceeded without undue sacrifice in mission accomplishment and an appropriate recovery can be made. In the present instance it seems evident that within some limits most of the typical criteria used could be exceeded without mission failure occurring—except in the narrow sense of not meeting all criteria. For example, a ΔV error of 11 fps is not appreciably more significant than a ΔV error of 9 fps, and so forth. Consequently, reliabilities with respect to such arbitrary criteria must be taken with due regard to their actual implications for the mission as well as their face value. Probabilities of mission success rather lower than ordinarily expected may not be worthy of serious concern after all.[5]

Thus, in keeping with the ambiguities involved in specifying reliability by means of arbitrary criteria and the belief that the pilots of C-8 and C-13 could have achieved still greater reliability, the present data should not occasion doubts about pilot capability to satisfactorily perform the type of mission used in this study. With carefully arranged additional training the performance reliabilities attained would likely exceed those calculated from retraining performance in this study and on close scrutiny they would be found to represent quite acceptable performance. However, along with careful attention to the training, periodic assessment of performance reliabilities would be necessary to assure this happy result—either in a simulated or in a real mission. The attenuation of mission effectiveness as missions become more lengthy and involved necessitates more than a casual, offhand approach to training and a quantitative validation rather than an informal judgment of performance capability.

---

5 This argument is also applicable to the concern for skill retention and so reduces the severity of the problem. But it does not eliminate the problem completely or the need for special attention to it in system design.

Methodology: The Measurement/Prediction of Human Performance

In addition to these specific implications for space system design, this study (more particularly the analysis presented) also illustrates some novel methods for handling human performance data. Because this is the first formal application of these methods a few general comments about their utility—both for the measurement and/or prediction of human reliability in system development and operational planning and for the quantification of behavior generally—seem in order.

Human Reliability in Development and Operational Planning. First, the outcome of the analysis has a familiar ring. Not only are the reliabilities obtained surprisingly compatible with general impressions of pilot capability but that expression of capability is familiar to systems engineers. For, in anticipating the effectiveness of systems under design and development, reliability studies estimating mean time between failure and, in turn, likelihood of successful operation in the single case are required. Only when such information is available can appropriate cost-effectiveness trade-offs be made and maximum functional value be obtained from the system. Furthermore, subsequent operational planning must or should be prefaced upon detailed information on specific capabilities.

However, even though this approach to system development has been widely discussed and is now well-accepted the full implementation of it with respect to all critical performances within the system has not and is not now being accomplished. As has long been recognized by human factors specialists and others, the capabilities of the people involved (either as operators or servicers) have a considerable influence on effectiveness (see, for example, Shapero et al, 1960 and Thomas, 1962). Yet studies of system effectiveness have continued to discount the element of human performance in favor of preoccupation with the more familiar and understandable machine elements. Machine reliability is required for system effectiveness—but so is human reliability in most systems of interest.

Among the several reasons for this persistent disregard of human performance in system design the reason of particular relevance here (and perhaps most important of all) is simply that human performance data, when available, are usually not in a form amenable to direct use in effectiveness studies. In predicting on a particular system what one needs to know is what are the probable limits of this individual's (or this select group of individuals') performance of these particular activities. The average or most typical performance of a general sampling of individuals on a collection of traditional laboratory tasks is not of immediate use and seemingly foreign to the design problem. Furthermore, the necessary interpreter of such data is often not available. In contrast, the analytic methods of this study offer the special merit of placing these human performance data in terms of immediate utility and that is why the results obtained seem familiar (to the systems engineer!).

This gratifying result is primarily the consequence of the fundamental innovation of interpreting observed performance probabilistically. After all, human performance is variable (to the great dismay of psychologists who have sought to express invariant relations involving it) and, therefore, statements about it are probabilistic, whether the fact is recognized or not. To express performance, then, directly as a probability is merely to explicitly recognize the variability which does exist. To do so is to stay desirably close to observational realities and avoid preoccupation with finding supposed "true" values, and so forth.

Now the expression of performance as a probability requires that some reference be chosen—i.e., that the probability be with respect to something. Two main alternatives are possible and they are both exemplified in the analysis of this study data. One may either refer to some absolute criterion dictated by outside considerations (system requirements, etc.) or one may refer to some relative criterion established by the individual's own or some other individual's performance. The first option is straightforwardly an application of the probability concept which relates human performance to the requirements for it in such a manner as to directly express capability of meeting those requirements. Consequently, it is extremely useful in making design choices as well as assessments later of operational capability.

On the other hand, the second option provides an extremely sensitive method of indicating change or difference consequent upon a change in conditions or personnel. One simply chooses an appropriate probability and establishes that level or standard of performance in the units of measure for the reference. The probability estimated from any other performance of achieving that standard level may then be directly compared with the reference probability and the amount (absolute or relative) of shift in probability taken to indicate the magnitude of the effects. In this way performance change is described probabilistically rather than in terms of the arbitrary units of measure (such as feet per sec., degrees, etc.). This option has special value for empirical determination of the relative merits of two or more design arrangements when outside criteria either do not exist or do not readily discriminate among the arrangements. But in many cases the information provided by reference to both absolute and relative criteria is needed—as was true in the present study.

In expressing relative reliability it is necessary to choose a reference probability and so it is naturally of concern what choice is best or how the choice should be made. In this instance $p = .950$ was chosen, but why was this value rather than, say, $p = .500$ (the mean of a normal distribution) selected? The most reasonable answer is that there is no one best value for all circumstances but rather a need to choose the value on the basis of the immediate purpose. In the present case, the interest was in performance levels attainable at a high level of consistency and so the .950 was selected. Furthermore, it is usually true that what one is interested in is the performance achievable with near certainty (or else that almost certainly not achievable) rather than merely the most typical performance. This suggests that a reference in the order of $p = .950$ would not generally be amiss. But, circumstances dictating an interest in the most typical ($p = .500$) or in some intermediate value (such as $p = .750$) can readily be imagined. Whatever the choice, the essential point to be emphasized here is that in describing performance probabilistically this significant choice for any kind of statistical analysis is made explicit. In this there is great merit because the traditional practice of describing performance in terms of central tendency—discounting variability as it does—deserves justification each time it is used.

Finally, in depicting performance as a probability a very considerable advantage is realized in the handling of realistic behavioral complexities such as typically are found in operational systems. One of the great faults of conventional psychological methods is they do not provide a satisfactory means of describing complex performances which are characterized by two or more measures (much less 22, as in this study). Yet, in reality, it is an inclusive expression of the totality of a performance which is most often desired. Even the normaliza-

tion and subsequent combination (weighted or not) of the several data series does not remedy this difficulty, for what results is merely an index relative to the particular data collection circumstances and having no independent meaning. In contrast, the expression of performance as a probability not only similarly permits a combination of the several measured aspects to arrive at a joint indicant but provides statements of probability to which other events may be related (as within a system). The result is that analysis in terms of probabilities allows, where conventional methods do not, the modeling of a complex series of man-machine functions necessary to effectiveness studies.

Thus, it is concluded that the analytic methods employed in this study also constitute valuable techniques for introducing human performance considerations into the design of systems. By describing performance probabilistically the advantages of staying close to the realities of the data and of using terms familiar to many design personnel are realized. Concern for human performance is encouraged, thereby, as it should be. By interpreting performances as the likelihood of meeting absolute criteria, derived from system analysis, the opportunity to influence specific design choices as well as to assess operational capability with respect to human performance is realized. For example, simulation tests of actual skill retention requirements can be properly evaluated and the effectiveness of proposed remedial techniques, when required, can be checked. By interpreting performance as the likelihood of meeting relative criteria, derived from the probabilistic description of another performance, an extremely sensitive indication of differences or changes in performance is obtained. This is a powerful technique for exploring particular design issues and guiding operational plans concerning training and other details. For example, a simple extension, such as the training decision rule alluded to earlier, can be used as the basis for realistic quantitative evaluation of achievement in training, thus eliminating personal judgments as the primary basis for training decisions. This would assure the desired human reliability in complex and lengthy missions by more precise determination of day to day training needs and, in turn, more effective training. Finally, by expressing human performance in various aspects of the job probabilistically the desirable option of formulating statements of overall probability and of entering human performance in system effectiveness models is gained. These are some of the more significant and obvious merits of the analytic techniques used—but not necessarily all their merits.

But what about the other side of the ledger, one may ask. What are the drawbacks and disadvantages of these methods? From a practical point of view it should be noted that a great number of calculations are required—a tedious process if accomplished by hand (as for this study). But, in any settled application—say in conjunction with mission simulation—computerized data reduction would be simple to arrange and of trivial cost. The calculations required are simple ones and in themselves rather familiar, at least when the normal probability model is adopted. The novelty involved is in the adaptation and combination of these basic and familiar techniques. On-line calculation, as might be of real advantage in a training situation, is imminently practicable. Furthermore, although the full-blown methods and associated jargon initially would seem strange to human factors specialists trained in traditional methods, their natural desire to see human performance more fully considered and the familiarity of the basic concepts should dispose them to quickly grasp the methods and apply them, as required. In fact, judging from present trends, the only system design personnel who probably would not readily accept these methods are the operational personnel (i.e., the research pilots involved) who are

generally disinclined to have numbers placed upon their performance. Even this may not be too serious a drawback, for if a suitable trial usage is arranged and due consideration is given to pilot suggestions as to details (measured parameters, models, etc.) it is possible that many such individuals would become more accepting. For when used in the proper way these methods avoid, at least, some of the deficiencies in conventional measurement methods which have prompted, justifiably, some of the misgivings of operational personnel about performance measurement. But, convinced or not, the quarrel of operational personnel is likely to be with measurement per se rather than with these methods and that, of course, is more of a political than a technical issue.

In short, there seem to be a variety of advantages and virtually no disadvantages attendant upon applying the analytic methods used in this study in system development and operational planning.

Quantification of Behavior. But what about the still broader implications of the analysis for the quantification of behavior, apart from the concern for particular systems? If the methods are useful for the quantification/prediction of behavior in man-machine systems, then perhaps they also ought to be useful in quantifying behavior in other circumstances, as in laboratory studies having no intended specific application. And, indeed, it is believed that this expectation is entirely warranted. Although the arguments are not so evident why psychological research would also benefit greatly from considered usage of these methods (and this is not the appropriate place to detail them) they parallel to a great extent with elaborations and ramifications those already given concerning systems applications. Certainly it would be beneficial and desirable to place research data in a form amenable to direct generalization (if permissible otherwise) to specific cases if this can be accomplished without sacrifice in other research goals. With these methods such is possible and in this way the bridge between the laboratory and the world of applications can be strengthened. But, apart from this very real advantage in the application of behavioral research there are also advantages to be gained in the way of improved, more effective research. A more complete statement of this view on analytic methodology and the justification for it may be found in another report (Cotterman, 1967).

Further Research

Finally, in addition to the implications for space system design and the measurement/prediction of human performance the results obtained suggest directions which the course of further research should take. There are many research possibilities which could be mentioned, but only the seemingly more urgent will be considered in the following.

Skill Retention. With respect to skill retention in complex and lengthy space missions it certainly cannot be asserted that this study provides all the needed answers. The significance of the concern for skill retention is believed established by this study but this is not enough. What is needed, in addition, is refined information on exactly how much loss is to be expected as a function of time periods ranging from a week or two to the limit of the longest mission contemplated in the foreseeable future. Furthermore, this relationship should be established for a sufficient number of discriminably different kinds of tasks as to be representative of the anticipated behavioral requirements, for skill retention is known to depend upon the type of task involved. To accomplish this

will require utmost care in research to establish, among other things, that prior to test the personnel involved have achieved known and equal levels of achievement. The training decision rule referred to, makes this possible, although some further development of it before application is desirable. Such information must be available if accurate prediction on the magnitude of the skill retention problem ever is to be made during the course of design. Without it the only recourse is to an empirical determination via simulation in each and every instance.

In addition, prediction of the magnitude of the skill retention problem is still not enough. Again, unless a quick fix in each instance is an acceptable strategy, ways of ameliorating the problem should also be evaluated, developed, and refined relative to the range of behavioral requirements of interest. Research having these goals might range all the way from specific changes in task requirements, through various training manipulations either originally or during the intervening period, to consideration of relevant selection factors. Even if it were limited to training remedies a programmatic effort likely to make serious headway with this problem would have to be extensive. Obviously, much more research on the skill retention problem is needed.

Measurement Methodology. Similarly, although the analytic techniques employed in this study are evidently considered to represent a sizeable step forward in capability to introduce human performance measurement in system design, development, and analysis and a contribution to behavior quantification generally, they also are not complete. Among the more significant needs in this regard are the following: (1) development of applications with the use of other than the normal probability (or the t) model, (2) adaptations and confirmation of tests of significance for application to probabilities in various circumstances, (3) more intensive analysis and development of the fundamental implications and rules for application, and (4) development of more refined methods of modeling complex events (i.e., events having multiple measured aspects which may be at least partially interdependent). As further work of this nature is accomplished and made available, the scope and utility of this novel measurement methodology will be greatly enhanced, with benefits both to psychological applications and to behavioral research.

SUMMARY

The primary intent of this study was to obtain a more valid estimate than was otherwise available of the degree of skill loss which may be expected over time intervals up to three months duration in tasks typical of space vehicle operation. To accomplish this, four crews of three each aerospace research pilots who had trained for a period of six weeks in the performance of a simulated 7-day lunar landing mission were tested various periods of time after the conclusion of training. Each crew was tested after a different interval—the intervals being approximately 4, 8, 9, and 13 weeks. The training had culminated in the real-time performance of the 7-day mission; however, for test purposes this mission was compressed into a single 13-hour workday by the omission of long coast phases, navigational tasks, certain secondary activities, and (for two crews) the duplicative transearth insertion phase. By this economization of time and costs, one day of additional training for the 4- and 9-week crews and three days of additional training for the 8- and 13-week crews was provided as a means of checking upon how quickly former skill in selected phases might be reacquired, or possibly surpassed. In the test mission and in the retraining all the pilots in each crew performed in all positions (pilot, navigator, and engineer) so that data were obtained from each of the twelve. But, with the exception of transposition by 4- and 9-week crews, the commander was always tested in the pilot position first so that one record of critical test mission performance from each crew would be uncontaminated by prior participation. Extensive performance records of flight control parameters and switching activities were obtained.

The analysis of the performance records focused attention upon 22 flight control parameters considered critical in the performance of the nine main phases of the test mission. The phases reflected by these are, successively, translunar insertion, transposition (of modules), lunar orbit insertion, separation and deorbit, braking and hover (lunar landing), lunar powered ascent, rendezvous, docking, and earth entry. The number of parameters considered descriptive of any given phase ranged from one (for two phases) to four (for one phase). The switching data are summarized for reference but because of their nature they are discounted as contributing little in this study to the understanding of skill retention.

By the use of novel analytic methods, extensive analyses of the flight control performance at the end of training, in the test mission, and in retraining were accomplished in such a way that the level of performance observed is given as a probability (or reliability value) rather than in the units of measure. In the main analysis the probabilities make reference to hypothetical system criteria for the 22 parameters and, thus, indicate the likelihoods of meeting these criteria. In a secondary, parallel analysis of test and retraining performance the probabilities make reference to the level of performance estimated to be achievable by each individual in 95% of his performances under like circumstances, thus, indicating the likelihoods of meeting these former levels of highly reliable performance. In both analyses the probabilities of meeting individual parameters within given phases are taken cumulatively ($\bar{p}$) as an estimate of likelihood of success in any parameter and jointly ($\dot{p}$) as an estimate of likelihood of success in all parameters of the phase. Similarly, these phase probabilities are taken cumulatively as estimates of success in any parameter ($\bar{\bar{p}}$) or as estimates of success in any phase ($\bar{\dot{p}}$) in the whole mission (or a collection of phases as in retraining). The probabilities of success in all parameters of

a phase also are taken jointly ($\dot{p}$) as estimates of success in all parameters and phases of the whole mission (or a collection of phases)—i.e., the probability of mission success. In this way overall performance indicants for each pilot were obtained.

On the basis of careful study and cross-comparison of the probabilities for individuals and (taken cumulatively) for crew performance as well as the nature and extent of possible biases involved, a number of specific findings were expressed. With regard to original capability at the end of training the crews tested after 8 and 13 weeks were found to have achieved quite high reliabilities ranging for individuals from .932 to .990 in meeting any criterion ($\bar{\bar{p}}$) and from .294 to .817 in meeting all criteria ($\dot{p}$). However, the lower two sets of values are considered underestimates due to the inclusion of unrepresentative scores, whereas all others are considered overestimates of true capability in a real mission of this nature (to the extent of about .005 in $\bar{\bar{p}}$ and .10 in $\dot{p}$). Hence, their typical likelihood at that time of meeting any criterion in a real mission is judged to have been around .975 while their likelihood of meeting all criteria is judged to have been around .615. In contrast, the crews tested after 4 and 9 weeks were found to have had much less capability at the end of training ($\bar{\bar{p}}$ and $\dot{p}$ never greater than .850 and .002, respectively). Because these crews obviously were not fully trained their data were discounted as incapable of supporting useful generalizations about skill retention in space missions (although analyses of them are included).

In the test mission, performed approximately 8 or 13 weeks after training, the well-trained crews showed an estimated capability to meet the criteria somewhat lower than their original capability to meet them. The first-tested pilots evidenced considerable loss with estimated $\bar{\bar{p}}$ values dropping to .931 and .899 and $\dot{p}$ values dropping to .048 and <.001, for the 8- and 13-week pilot, respectively, but, some of those tested second and third showed a slight (in part, spurious) gain—the crew averages, consequently indicating mild losses. As was expected, the same pilots showed even less likelihood of reaching former highly reliable ($p_{.95}$) levels of performance in individual parameters. Evidently, lack of direct practice over a period of 8 weeks or more resulted in a definite loss in skill which was to a considerable extent recovered by the opportunity to observe or participate once or twice in another pilot's performance of the mission. However, a difference in retention over 8 and 13 weeks was not demonstrated—perhaps because of differences in original learning—which indicates that such variations in retention interval may be of little significance when the intervals are that long. The levels of reliability derived from the first test performance of around .92 of meeting the criterion in any parameter and of around .02 of meeting the criteria in all parameters are considered to somewhat overestimate real mission performances under like circumstances of previous training, intervening activities, etc.

In the course of three days of additional training the well-trained pilots (those tested after 8 and 13 weeks) quickly regained and eventually surpassed their original capability to meet the system criteria. Their typical probabilities of successful performance in a real mission at this time are estimated to have been about .996 for $\bar{\bar{p}}$ and .936 for $\dot{p}$; this implies that they had not quite reached the limits of their potential in original training. But, even at their best, these pilots did not demonstrate in retraining their full original capability (relative to former $p_{.95}$ levels) to perform all the measured aspects of flight control. Their efforts to maximize performance with respect to criteria

with accompanying frequent shifts between training and retraining in relative performance on aspects within phases, along with the sensitivity of this analysis to such shifts, is considered the probable explanation. Such changes in strategy, which are costly of overall reliability, are not reflected in the criterion-reference analysis because they are generally beyond the arbitrary maximum discriminable reliability of .999.

These specific findings are considered to have a number of valuable implications for space system design and for the measurement and prediction of human performance. Thus, it is concluded with respect to skill retention in long space missions that a requirement to remember a critical flight control task 8 weeks or more does pose a problem worthy of special design attention. However, assuming the intervening activities maintain alertness to time-sharing and rapid response and introduce no interference, up to at least 13 weeks the problem is not likely to be so severe that it cannot be remedied by judicious use of training techniques prior to critical task performances for refamiliarization—particularly with timing and time-sharing aspects. More refined information than this study provides is needed to allow precise prediction about skill retention problems and planned remediation in design. With respect to pilot reliability in space missions it is concluded that on the basis of these data (maximal capability in retraining) no doubts about pilot capability to perform the type of mission used in this study are justified providing extreme care is given to training and to obtaining demonstrated performance reliability in the course of it.

With regard to the measurement and prediction of human performance it is contended that the novel analytic methods used constitute valuable (and needed) techniques for introducing human performance considerations in the design of systems. They appear to offer many advantages and almost no disadvantages in this application. Furthermore, they are also viewed as offering a large potential use in the quantification/prediction of behavior generally. Accordingly, as continuing efforts to enhance the scope and utility of this novel measurement methodology are accomplished considerable benefit to both psychological applications and to behavioral research are to be expected.

APPENDIX I

Summary of Switching Data

TABLE XXIV

Skill Retention Test Mission Switching Errors By Phase and Workspace

Crew/ Position/ Pilot			Phase *												
								Lunar Landing				Lunar Ascent			
			EA▼	TLI▼	TRN	TRN/ CHK	LOI	SDO	CD	BH	LPA	CA	RD	EE
C-4/P	SW		142	21	17†	222	22		57			25		39
	1		9			NP						2		2
	2		NP		1(19)**	NP						2		1
	3		NP			4	(24)		2			2		
E	SW		258	6†	4	NP	10		111			110		17
	1	(P-2)	3				3		3			3		
	2	(P-3)	NP				(11)					2(114)		
	3	(P-1)	NP		1(12)		(16)		1			3(118)		1
C-9/P	SW		143	21	17†	222	22		57			25		39
	1					NP						1		
	2		NP	1		NP			1			1		
	3		NP		(19)	5	(24)					2		
E	SW		260	6†	6	0	16†		112			117		17†
	1	(P-2)	8				1		7			5		1
	2	(P-3)	NP		1		(11)		(111)			25(108)		
	3	(P-1)	NP				(11)		6(111)			6(114)		1
C-8/P	SW		130	24	19	27	20	24	7	10	15	2	4	39
	1		2							1				1
	2		NP			NP								
	3		NP			NP								
E	SW		223	10	20	160	16	58	47	26	98	13	21	24
	1	(P-3)	3											
	2	(P-1)	NP			NP	1		(44)	1			(20)	
	3	(P-2)	NP			NP				1	1			
C-13/P	SW		130	24	19	27	20	24	10	15	17	6	5	39
	1													
	2		NP			NP								
	3		NP			NP								
E	SW		220	10	20	160	16	58	44	21	96	9	20	24
	1	(P-3)			1		?							4
	2	(P-1)	NP			NP		(15)			1			1
	3	(P-2)	NP			NP			1		1			

* Lunar Landing and Lunar Ascent data were broken into lesser elements only for C-8 and C-13. CD is Coast Descent and CA is Coast Ascent.

▼ in EA of C-4 the Navigator (P-3) had 20 switches, 1 error; C-9 (P-3) 20 switches, 1 error; C-8 (P-2) 53 switches, no errors; C-13 (P-2) 52 switches; 1 error. In TLI for C-4 and C-9 the navigator had 4 switches. Pilot performance order was P-2, P-3, P-1. P-2 of C-9 missed 1 switch.

† the order of performance in these instances was P-3, P-1, P-2.

** parenthetical numbers indicate switches expected when different from above.

NP not performed by this Pilot.

APPENDIX II

Notes on Analytic Methodology

APPENDIX II. NOTES ON ANALYTIC METHODOLOGY

In preparing the main text of this report the intention was to present enough general information on the analytic methods used to make the results meaningful and interpretable. It did not seem wise to make fully explicit all the initial computations, such as might be of interest to one wishing to verify a portion of the results or to adopt the methods to his own needs. Yet, being quite different from traditional methods in orientation and intent, the methods used should be fully described. Consequently, this appendix provides additional step-by-step details on the computations involved in obtaining the basic analytic results, from which the results described in the text were computed. More general discussion of the underlying measurement philosophy may be found in a separate report by Cotterman (1967).

The Data Base

As already described (see pages 24-27) the analytic problem began with a voluminous collection of flight control data for each of 12 pilots. By careful elimination this volume was reduced to sets of data on 22 parameters scattered throughout the simulated lunar landing mission, and further categorized by association with one of a number of phases. Most of these 22 measures represented direct records of flight control information, although four of them represented simple combinations of two measures (see page 25) which were treated in the same fashion as the raw data. Furthermore, as depicted in appendix IV, for each pilot with respect to each phase there were records of performance on first a variable number of training trials, then a real time mission, then the fast time skill retention test mission, and finally on a variable number of retraining trials on certain mission phases.

Computation of Parameter Probabilities (p)

The first step in interpreting this large collection of data was to arrive at a probabilistic indication of each pilot's performance capability in each measured aspect in training, in the skill retention test mission, and in retraining. The real time mission was omitted from consideration because a presumably more adequate indication of skill achievement was available from the multiple performances just prior to it at the end of training. Capability was assessed in two ways—first with reference to hypothetical system criteria for the measured aspects and second with reference to capability demonstrated by the individual in training (his $p_{.95}$ level per measured aspect).

Reliability per criteria in training. Calculation of performance capability or reliability at the end of training with regard to the hypothetical criteria involved three simple, though tedious, steps. First, having chosen to consider the last four training trials (with the few exceptions noted in the text) as best representing that capability, means and estimated standard deviations were computed from each sample. The formulas used are the conventional ones which may be given as follows:

(1) Mean = $\bar{X} = \frac{\Sigma X}{n}$

(2) Estimated Standard Deviation = $s = \sqrt{\frac{n\Sigma X^2 - (\Sigma X)^2}{n(n-1)}}$

in which X is the recorded datum and n is the number of such data items (4 in this case).

Next, on the assumption that repeated such sets of four performances under exactly the same circumstances would be found to follow the normal probability density function the relevant criteria were interpreted in terms of a normal function having these respective means and estimated standard deviations. Thus, by the familiar operation of dividing the difference between the mean and the criterion by the estimated standard deviation the distance (z), in standard deviation units, of the criterion from the sample mean was obtained in each instance. This relationship may be summarized as:

(3) Standard Deviations from the Mean = $z = \frac{\bar{X} - C}{s}$

in which C is the criterion for a particular measured aspect. (See table V on page 29 for a listing of the hypothetical system criteria.)

The z values so obtained also represent the number of standard deviations of the criterion from the mean of a normal distribution having a mean of zero and a standard deviation of one, for which the proportions above and below a particular deviation are well known. Accordingly, in the final step, the z-value obtained for each measured aspect and each pilot was interpreted as a probability by simple interpolation from a standard table of the normal probability integral (Peters and Van Voorhis, 1940, pages 481-4). The probability taken was always the proportion of the hypothetical function (based on the sample statistics) indicating performance superior to the criterion. In actuality, this was performance numerically smaller than the criterion.

Thus, for example, to determine the capability of P-131 to meet the criterion of a 10 fps rate at impact on lunar landing the recorded rates on the last four training trials were taken as the sample. These rates were 2.8, 5.2, 5.6 and 5.8 fps. The $\bar{X}$ of these values was found to be 4.85 and the s was found to be 1.3892, as computed by formulas 1 and 2. Also, by formula 3, the difference of 5.15 between criterion and mean was found to give a z of 3.7072. Or, described in another way, the mean of the sample was found to be nearly four standard deviations superior to (though numerically less than) criterion performance. According to the tabled values of z for the normal probability integral, somewhat fewer than 5 cases in 10,000 would be expected to fall beyond a z of 3.7072, and thus to exceed the criterion, whereas the remaining 9,995 or more cases would be expected to be less than the criterion. Therefore, the probability of P-131 meeting the criterion

of a 10 fps rate at impact on lunar landing is >.999, as given in table XXV (appendix III). This circumstance is illustrated in figure 10, in which the hypothetical distribution of P-131's performance at the end of training is shown with relevant values noted.

Reliability per criteria in test. The assessment of capability in the skill retention test followed the same logic, but with variations in calculation necessitated by the availability of only one measurement. If the test of each pilot's skill had involved multiple performances then it would have been possible to take some (or all) of them as a sample and to proceed exactly as with the training data to arrive at the likelihood of success in meeting each criterion. However, in order to avoid inaccuracy because of relearning only the first performance after the retention period could be considered and, with only one measurement, neither the mean nor the estimated standard deviation could be computed.

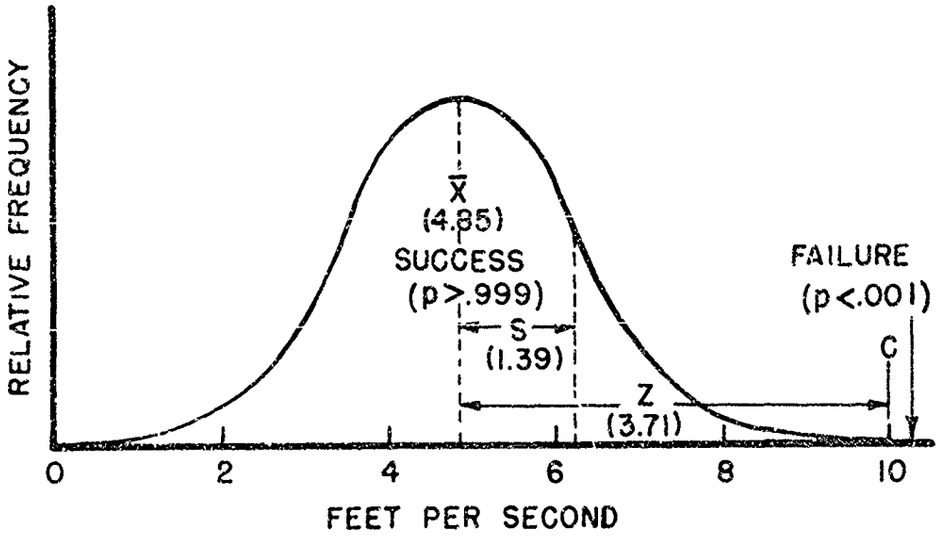


Figure 10. Hypothetical Distribution at the End of Training of P-131's Vertical Rate at Impact on Lunar Landing indicating Probability of Meeting the Criterion

This analytic problem was resolved by assuming that the obtained value constituted the best available estimate of the mean and that the variability in performance at the end of training is indicative of what would be observed in multiple test performances under the same circumstances (if they were possible). That variability is to be expected implies that the single performance would not be the same as an observed mean. Also, if performance change between training and test is to be expected then the variability in performance may be expected to change in some degree. Consequently, some inaccuracies are to be expected in the use of these assumptions. Nevertheless, the estimates of capability so obtained are based on the best available information and are probably biased as a group to only a slight extent, as discussed on page 76.

Implementation of these assumptions in calculating the estimated probabilities was comparatively simple. With the test performance (X_M) and s for training already available, to arrive at the standard deviations between test and criterion performance (z_M) it was merely necessary to divide the actual difference between them by the borrowed standard deviation estimate, as in the following:

$$z_M = \frac{X_M - C}{s_T} \tag{4}$$

in which X_M is the recorded mission performance and s_T is the estimated standard deviation of performance in training on a particular measured aspect. Again, the proportion of values expected to be numerically less (hence superior) to a z as great as that obtained was taken as the probability of successfully meeting the criterion.

By way of further illustration, P-131's lunar landing impact rate of 5.4 fps in the test mission is 4.6 fps less than (or superior to) the criterion. This difference is 3.3113 times the standard deviation of 1.3892 fps found in training, suggesting that slightly more than 9,995 such test performances in 10,000 distributed normally with a mean of 5.4 fps and a standard deviation of 1.3892 fps would not exceed the criterion of 10 fps. Consequently, P-131's reliability in controlling the lunar landing impact rate in the retention test was considered to be >.999, as shown in table XXVI (appendix III).

Reliability per criteria in retraining. The assessment of capability to meet criteria in retraining required new calculations but, unlike that for test performance, presented no complications because multiple measures were available. Thus, in general, reliabilities per criteria in retraining were computed in exactly the same way as reliabilities per criteria at the end of training. A sample of performances was identified and $\bar{X}$ and s computed from that sample. Then, z with respect to the criterion was calculated and interpreted as the probability (p) of successful performance. Of course, as described in the text, for the six pilots comprising two crews (C-8 and C-13) this was done both with reference to the first four and the best four trials of retraining (taken in successive four-trial blocks). Calculation of the reliabilities of the six pilots comprising C-4 and C-9 who were tested earlier was based upon all their retraining trials as a sample because they had fewer trials. But these are differences in sampling, not in method of calculation.

Probability of achieving former levels of highly reliable performance. The second assessment of capability, rather than referring to the hypothetical system criteria, referred to a level of performance in each measured aspect which each individual had demonstrated capability to achieve with high reliability. Instead of an arbitrary and absolute criterion common for the analysis of all the pilots' performances, in this assessment the reference was to an individually-established level of performance varying with the individual and his own former skill. The value of this second assessment was its greater sensitivity to shifts in performance in the test mission and in retraining, from that observed at the end of training.

Calculation of these probabilities for each pilot on each measured aspect involved one new calculation and a simple manipulation of the result in conjunction with certain values already available from the analysis with respect to criteria. Having chosen a probability of .950 as a reasonable compromise between certainty of achievement and estimation error, the first step was to estimate for each measure and pilot the level of performance to be expected in 950 performances of 1,000 at the end of training (designated $\dot{X}_{.95}$). This was accomplished by reference once again to the normal distribution which is known to include 95% of the cases from one extreme to 1.6449 standard deviations in the direction of the other extreme. Thus, to find the $\dot{X}_{.95}$ reference it was merely necessary to add 1.6449 standard deviations to the mean, as in the following:

$$\dot{X}_{.95} = \bar{X} + 1.6449\, s \tag{5}$$

Subtraction instead of addition would have been required if the measures had increased rather than decreased with superiority in performance. The $\dot{X}_{.95}$ so obtained provided the reference performance level, analogous to a system criterion, to be used in this second assessment of capability in test and retraining.

With the $\dot{X}_{.95}$ available for each measured aspect and pilot it was then a simple, repetitious operation to estimate the probabilities of achievement. As in the analysis with regard to criteria, the difference between mean performance (or test performance taken as the mean) and the $\dot{X}_{.95}$ divided by the estimated standard deviation indicates the number of standard deviations of the $\dot{X}_{.95}$ from the mean. For the test performance this relationship is

$$z_{M/.95} = \frac{X_M - \dot{X}_{.95}}{s_T} \tag{6}$$

in which, as in formula 4, the s_T computed from training performance is used as the estimated variability. For retraining performance the relationship is

$$z_{R/.95} = \frac{\bar{X}_R - \dot{X}_{.95}}{s_R} \tag{7}$$

in which the subscript R indicates that the value was computed from the retraining sample. Both expressions (6 and 7) are the same as their counter-

parts for computing z with respect to criteria except that $\dot{X}_{.95}$ is substituted for C. The resulting z-values are likewise used to obtain, by interpolation from a table of the normal probability integral, the expected proportion of cases superior to (or less than) the reference—in this case $\dot{X}_{.95}$.

Thus, pursuing the previous illustrative data still further, P-131's impact rates on his last four lunar landings in training suggest (by application of formula 5) that in 95 out of 100 performances his rate would not exceed 7.1351 fps. This $\dot{X}_{.95}$ differs by 1.7351 fps or 1.249 estimated standard deviations (of 1.3892 fps) from his retention test rate of 5.4 fps. A deviation that great at one extreme of the normal distribution is expected to occur only about 106 times in 1,000. Consequently, the likelihood in the test mission of P-131 meeting his former $p_{.95}$ level of 7.1351 fps in rate at impact on lunar landing is taken to be .894. This outcome indicates a slight loss from training to test in capability to control impact rate (i.e., a decrease in likelihood of .056 from the reference of .950).

Furthermore, in his best four-trial block of retraining P-131's impact rates were 3.0, 1.4, 3.4, and 2.6 fps with a $\bar{X}_R$ of 2.6 and an s_R of .864 fps. The $\dot{X}_{.95}$ of 7.1351 computed from the training data is found to differ by 5.249 estimated standard deviations from the $\bar{X}_R$ of 2.6 and a deviation that great in the same direction is expected to occur considerably fewer than five times in 10,000 performances. The estimated likelihood in retraining of P-131 successfully realizing his former $p_{.95}$ level of impact rate control on lunar landing is, then, in excess of .9995. This represents a considerable increase in capability over that demonstrated at the end of training.

Summary of procedures for obtaining parameter probabilities (p). By the use of the simple methods just described and illustrated all the values tabled in appendix III were obtained. These values provide the desired probabilistic indication of each pilot's performance capability in each measured aspect of training, test mission, and retraining to serve as a basis for assessing, more inclusively, phase, mission, and crew capabilities. None of the computations required are unusual or complicated, but when there are many of them and they are accomplished manually (as in this instance) they do require considerable time. Therefore, when applied to data of much volume these methods would be most efficiently carried out by means of computer analysis. The following summary of procedural steps required in this initial analysis suggests the ease with which the analysis might have been programmed for digital computation.

1. Identify and note (enter) all reference performance consisting of:

   a. hypothetical system criteria (C)

   b. training performance sample (X_T)

   c. test mission performance (X_M)

   d. retraining performance samples (X_R)

parts for computing z with respect to criteria except that $\dot{X}_{.95}$ is substituted for C. The resulting z-values are likewise used to obtain, by interpolation from a table of the normal probability integral, the expected proportion of cases superior to (or less than) the reference—in this case $\dot{X}_{.95}$.

Thus, pursuing the previous illustrative data still further, P-131's impact rates on his last four lunar landings in training suggest (by application of formula 5) that in 95 out of 100 performances his rate would not exceed 7.1351 fps. This $\dot{X}_{.95}$ differs by 1.7351 fps or 1.249 estimated standard deviations (of 1.3892 fps) from his retention test rate of 5.4 fps. A deviation that great at one extreme of the normal distribution is expected to occur only about 106 times in 1,000. Consequently, the likelihood in the test mission of P-131 meeting his former $p_{.95}$ level of 7.1351 fps in rate at impact on lunar landing is taken to be .894. This outcome indicates a slight loss from training to test in capability to control impact rate (i.e., a decrease in likelihood of .056 from the reference of .950).

Furthermore, in his best four-trial block of retraining P-131's impact rates were 3.0, 1.4, 3.4, and 2.6 fps with a $\bar{X}_R$ of 2.6 and an s_R of .864 fps. The $\dot{X}_{.95}$ of 7.1351 computed from the training data is found to differ by 5.249 estimated standard deviations from the $\bar{X}_R$ of 2.6 and a deviation that great in the same direction is expected to occur considerably fewer than five times in 10,000 performances. The estimated likelihood in retraining of P-131 successfully realizing his former $p_{.95}$ level of impact rate control on lunar landing is, then, in excess of .9995. This represents a considerable increase in capability over that demonstrated at the end of training.

Summary of procedures for obtaining parameter probabilities (p). By the use of the simple methods just described and illustrated all the values tabled in appendix III were obtained. These values provide the desired probabilistic indication of each pilot's performance capability in each measured aspect of training, test mission, and retraining to serve as a basis for assessing, more inclusively, phase, mission, and crew capabilities. None of the computations required are unusual or complicated, but when there are many of them and they are accomplished manually (as in this instance) they do require considerable time. Therefore, when applied to data of much volume these methods would be most efficiently carried out by means of computer analysis. The following summary of procedural steps required in this initial analysis suggests the ease with which the analysis might have been programmed for digital computation.

1. Identify and note (enter) all reference performance consisting of:

   a. hypothetical system criteria (C)

   b. training performance sample (X_T)

   c. test mission performance (X_M)

   d. retraining performance samples (X_R)

2. Compute $\bar{X}$, s, and $\dot{X}_{.95}$ for

   a. training performance sample

   b. retraining performance samples ($\dot{X}_{.95}$ not used)

3. Compute all z_c for

   a. training as $z_{c/T} = \dfrac{\bar{X}_T - C}{s_T}$

   b. test mission as $z_{c/M} = \dfrac{X_M - C}{s_T}$

   c. retraining as $z_{c/R} = \dfrac{\bar{X}_R - C}{s_R}$

4. Compute all $z_{.95}$ for

   a. test mission as $z_{.95/M} = \dfrac{X_M - \dot{X}_{.95/T}}{s_T}$

   b. retraining as $z_{.95/R} = \dfrac{\bar{X}_R - \dot{X}_{.95/T}}{s_R}$

5. Interpret all z as single-tailed probabilities (p)

Computation of Phase, Mission, and Crew Probabilities

With the probabilities describing each pilot's performance capability in each measured aspect of training, test mission, and retraining available it was a simple matter to obtain the various more inclusive probabilities reported in the fashion described on pages 27-28 and 33. Only two types of calculations were involved—although some of the indicants obtained (for example, $\bar{\dot{p}}$) required both kinds. In the one operation, the results of which are always designated by a bar, the sum of the values was simply divided by their number to obtain an average or cumulative probability for the set. In the other operation, the results of which are always designated by a dot, the joint product of the values was taken. The order in which these calculations were performed to obtain any given indicant is always shown by the upward progression of these bar and dot superscripts from the basic symbol. For example, in obtaining $\bar{\dot{p}}$, the joint products of the separate probabilities within phases were calculated first and then these products were taken cumulatively. With an awareness of these coding procedures all of the probabilities reported in the text may be interpreted readily.

APPENDIX III

Probabilities for Each Measured Parameter

TABLE XXV

Reliability Per Criteria at the End of Training (Last Four Trials)

Phase/Measure		Pilot 41	42	43	91	92	93	81	82	83	131	132	133
TLI	V	.785	--†	.929	.632	.625	.500	.979	--	.908	--	--	.806
	P	--	.914	.926	--	.442	--	--	--	--	--	--	--
TRN	DR	--	--	.987	--	--	--	--	--	--	--	--	--
	XR	--	.540	--	.336	.752	.503	--	.999	--	--	--	.999
	F	.234	.100	.016	.507	--	.570	--	--	--	--	--	--
LOI	V	--	.609	--	.576	--	--	--	--	.983	.999	--	--
	P	--	.986	--	.946	.992	--	--	--	--	--	--	--
SDO	V	--	--	.998	.472	.339	.868	--	.860	.999	.957	.658	--
	Y	.997	--	--	NA	NA	NA	.757	--	--	--	.565	.999
	P	.787	.480	.594	NA	NA	NA	--	--	.999	.840	.994	.985
B&H	D	.385	--	--	.437	.864	--	--	--	--	--	--	.993
	DR	.316	--	--	.364	.539	.969	--	--	--	--	--	--
	XR	.312	--	.992	--	--	--	--	--	--	--	--	--
	F	.786	.997	.989	.669	.848	.779	--	.999	--	.920	--	--
LPA	V	.358	.345	--	.158	NA	.170	--	.806	--	.999	.999	.557
Rend	F	NA	NA	NA	NA	NA	NA	--	--	--	.789	.978	.997
Dock	D	.935	.219	.600	.600	.935	.355	--	--	--	--	.995	--
	DR	--	.277	--	--	--	.479	--	--	--	--	.823	--
	XR	.244	.370	.653	.310	.235	.683	--	.988	.917	--	--	.827
EE	D	.539	.701	--	.758	.492	.269	--	--	--	--	.994	--
	A	.020	.308	.282	.813	.872	.439	--	--	--	--	--	--
	CR	.213	.597	.380	.286	.597	.022	--	--	--	--	--	--

† Dash indicates value exceeding .9995

NA Data not available

TABLE XXVI

Reliability Per Criteria in the Test Mission

Pilot

Phase/Measure		41	42	43	91	92	93	81	82	83	131	132	133
TLI	V	--†	--	.987	.928	.849	.803	.367	.500	.986	--	.993	.922
	P	--	.998	.996	--	.655	.993	--	--	--	--	--	--
TRN	DR	--	--	.999	<.001	--	--	--	--	--	--	--	--
	XR	--	.587	--	<.001	.075	.606	--	--	--	--	--	.996
	F	.255	.896	.116	.729	--	.884	--	--	--	--	--	--
LOI	V	--	.839	--	.896	--	--	--	--	.983	--	--	--
	P	--	.999	--	.997	.916	--	--	--	--	--	--	--
SDO	V	<.001	<.001	.959	NA	.517	.843	--	.939	.970	.889	.949	.940
	Y	.993	--	--	NA	NA	NA	.138	--	--	.999	.736	.999
	P	.601	.590	.518	NA	NA	NA	--	--	--	.991	.995	.988
B&H	D	<.001	--	--	<.001	<.001	NA	--	--	--	<.001	--	.925
	DR	.507	--	--	<.001	<.001	.999	--	--	--	<.001	--	--
	XR	.502	--	.847	<.001	<.001	.728	--	.998	--	--	--	--
	F	.390	.994	.597	--	--	.179	--	.873	--	.176	--	--
LPA	V	.635	.655	--	.583	NA	.557	--	.966	--	.903	.964	.716
Rend	F	NA	NA	NA	NA	NA	NA	--	--	--	.936	--	--
Dock	D	<.001	.672	.661	.611	.475	.473	--	--	--	--	--	--
	DR	<.001	.708	--	--	.936	.543	--	--	--	--	.975	--
	XR	.562	.583	.624	.308	.208	.312	.998	--	.946	--	--	.853
EE	D	.962	.872	--	.946	.557	NA	--	--	--	--	.993	--
	A	.895	.094	.085	.665	.914	.939	--	--	.997	--	.999	--
	CR	.354	.288	.054	.608	.908	.555	.942	--	.998	--	--	--

† Dash indicates value exceeding .9995

NA Necessary reference data not available

TABLE XXVII

Probability Estimated from Test Mission Performance of Attaining the $p_{.95}$ Level of the Last Four Training Trials

Phase/Measure		Pilot 41	42	43	91	92	93	81	82	83	131	132	133
TLI	V	--†	.998	.992	.997	.991	.994	.232	.001	.994	--	.424	.986
	P	--	--	.998	--	.986	<.001	.996	.902	.995	--	--	.282
TRN	DR	<.001	.940	.995	<.001	.991	.999	.992	.913	.988	.775	.992	--
	XR	.804	.961	.935	.004	.317	.972	--	.977	.909	<.001	.826	.893
	F	.956	--	.995	.987	.998	.996	.999	.999	.998	.997	--	--
LOI	V	--	.991	--	.997	.999	.996	.984	.663	.950	[illegible]	.999	.633
	P	.977	.994	.999	.997	.726	--	--	.991	.827	--	.932	.904
SDO	V	<.001	<.001	.670	NA	.982	.931	.397	.983	.671	.875	.998	.408
	Y	.906	.947	.811	NA	NA	NA	.445	.818	<.001	.686	.983	.926
	P	.866	.973	.927	NA	NA	NA	.760	.998	.987	.999	.955	.958
B&H	D	.001	.034	.989	<.001	<.001	NA	.508	.202	.848	<.001	.927	.735
	DR	.984	<.001	.550	<.001	<.001	.998	.691	.411	.981	<.001	.955	.216
	XR	.984	.900	.600	<.001	<.001	<.001	.696	.745	.955	.894	.999	.992
	F	.717	.918	.346	--	--	.484	.997	.366	.006	.245	.901	.291
LPA	V	.991	.993	--	.998	NA	.997	.933	.995	.998	.382	.657	.981
Rend	F	NA	NA	NA	NA	NA	NA	.969	--	.997	.991	--	.983
Dock	D	<.001	.998	.965	.953	.526	.974	.070	.997	<.001	.933	.995	.864
	DR	<.001	.997	.991	.998	.999	.965	.957	.941	.991	.998	.996	.831
	XR	.994	.986	.941	.949	.940	.949	.156	--	.959	.028	.998	.960
EE	D	--	.988	.756	.995	.965	NA	.189	.976	.011	.976	.940	.771
	A	--	.797	.802	.881	.969	--	.838	--	.797	.970	.028	.890
	CR	.981	.800	.634	.993	.997	--	.300	<.001	.751	.318	.963	.469

† Dash indicates value exceeding .9995
NA Necessary reference data not available

TABLE XXVIII

Reliability Per Criteria in Early Retraining **

Phase/Measure		Pilot 41	42	43	91	92	93	81	82	83	131	132	133
TRN(2)	DR	-†	-	.997	-	-	-	ND	-	-	-	-	-
	XR	-	.174	.987	.456	.452	.515	ND	-	-	-	-	-
	F	.455	.830	.830	.716	-	.781	ND	-	-	-	-	-
LOI	V	ND	ND	ND	ND	ND	ND	ND	-	-	-	-	-
	P	ND	ND	ND	ND	ND	ND	ND	-	-	-	-	-
SDO	V	ND	ND	ND	ND	ND	ND	.984	.971	-	-	-	.553
	Y	ND	ND	ND	ND	ND	ND	-	-	-	-	-	-
	P	ND	ND	ND	ND	ND	ND	.235	.998	-	-	.937	.987
B&H(8)	D	.326	.525	-	.357	.346	.377*	-	-	-	-	-	-
	DR	.389	.587	-	.450	.387	.411	-	-	-	-	-	-
	XR	.383	-	-	.418	.940	.772	-	-	-	-	-	-
	F	.672	.781	.627	.716	.793	.657	-	.991	-	.982	-	.811
Dock(6)	D	.688	.565	.958	.652	.851	.693	ND	-	-	-	-	-
	DR	-	.602	-	-	.827	.579	ND	-	-	-	-	-
	XR	.993	.627	.975	.542	.966	.711	ND	-	-	-	-	-
TEI(2)	V	.979	-	.991*	.775	.832	ND	ND	ND	ND	ND	ND	ND
	P	-	.404	- *	.758	.498	.930	ND	ND	ND	ND	ND	ND
EE(5)	D	.760	-	-	.940	.804	.648	-	-	-	.978	-	.432
	A	<.001	.062	.959	.347*	.785	.358	.615	-	.928	.999	-	.792
	CR	.404	.872	.997	.363*	-	.608	.462	-	-	-	-	.983

** For C_8 and C_{13} based on the first four training trials, for C_4 and C_9 based on all retraining data (generally 2 to 8 trials, depending on phase as noted in parentheses)

† Dash indicates value exceeding .9995

* Based on reduced data

ND Phase/parameter not done

TABLE XXIX

Probability Estimated from Early Retraining Performance of Attaining the $p_{.95}$ Skill Level of the Last Four Training Trials **

Phase/Measure		Pilot 41	42	43	91	92	93	81	82	83	131	132	133
TRN(2)	DR	.695	.988	.985	.988	.918	.878	ND	.901	-†	-	-	-
	XR	.778	.727	.078	.975	.800	.953	ND	-	.990	.997	.887	.996
	F	.988	-	-	.986	.998	.988	ND	.838	-	-	-	-
LOI	V	ND	ND	ND	ND	ND	ND	ND	.663	-	-	.972	.869
	P	ND	ND	ND	ND	ND	ND	ND	-	-	.771	.516	-
SDO	V	ND	ND	ND	ND	ND	ND	.701	.998	.981	-	-	.408
	Y	ND	ND	ND	ND	ND	ND	-	-	.022	-	-	-
	P	ND	ND	ND	ND	ND	ND	.041	.069	.998	-	.846	.959
B&H(8)	D	.406	.320	.939	.384	.354	.358*	.961	.173	.562	.758	.743	.990
	DR	.995	.340	.881	.927	.460	.406	.077	.924	-	.028	-	.999
	XR	-	.827	.996	.358	.702	.349	.595	-	-	-	.950	-
	F	.899	.591	.481	.844	.890	.813	.624	.837	.059	.991	.998	.506
Dock(6)	D	.713	.998	-	.990	.881	-	ND	.783	.860	-	-	-
	DR	.761	-	-	.782	.623	-	ND	.682	.773	.997	-	-
	XR	-	-	-	-	-	-	ND	-	-	.367	-	-
TEI(2)	V	.998	.989	.991*	.996	.986	ND	ND	ND	ND	ND	ND	ND
	P	.913	.022	.948*	.959	.998	.933	ND	ND	ND	ND	ND	ND
EE(5)	D	.978	-	.879	.998	-	.943	.521	.937	.127	.637	.983	.340
	A	.997	.939	-	.626*	.878	.912	.318	.999	.688	.568	.202	.305
	CR	.900	.999	-	.978*	-	.994	.309	.997	.990	.985	-	.244

** For C_8 and C_{13} based on the first four training trials, for C_4 and C_9 based on all retraining data (generally 2 to 8 trials, depending on phase as noted in parentheses)

† Dash indicates value exceeding .9995

* Based on reduced data

ND Phase/parameter not done

TABLE XXX

Reliability of C-8 and C-13 as Estimated from Best Four-Trial Block in Retraining

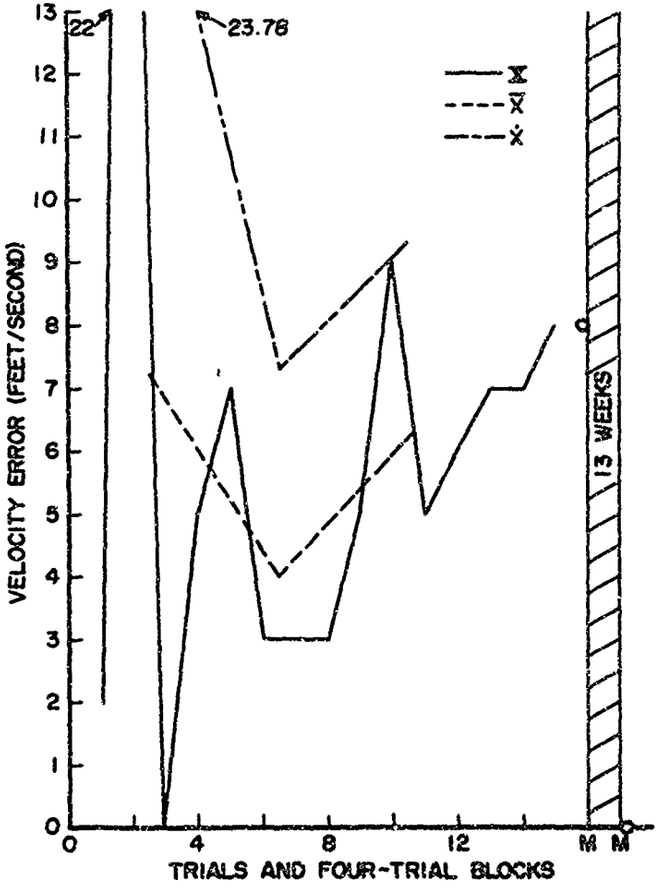
Pilot

Phase/Measure		Per Criteria						Per $P_{.95}$ Training Level					
		81	82	83	131	132	133	81	82	83	131	132	133
TRN	DR	NA	-†	-	-	-	-	NA	-	-	-	-	-
	XR	NA	-	-	-	-	-	NA	.999	-	.997	.887	.996
	F	NA	-	-	-	-	-	NA	.998	-	-	-	-
LOI	V	NA	-	-	-	-	-	NA	.663	-	-	.972	.999
	P	NA	-	-	-	-	-	NA	-	-	.771	-	-
SDO	V	.986	-	-	-	-	.973	.695	-	.262	-	-	.821
	Y	-	-	-	-	-	-	-	.997	.183	-	-	-
	P	-	-	-	-	.937	-	.961	.402	-	-	-	-
B&H	D	-	-	-	-	-	-	.979	.289	.990	.998	.811	.990
	DR	-	-	-	-	-	-	.216	.915	-	.535	.963	.999
	XR	-	-	-	-	-	-	.562	-	-	-	.969	-
	F	-	-	-	-	-	.811	.897	.656	.812	-	-	.506
Dock	D	NA	-	-	-	-	-	NA	.973	.953	-	-	-
	DR	NA	-	-	-	-	-	NA	.590	.988	.997	-	-
	XR	NA	-	-	-	-	-	NA	-	.819	.367	-	-
EE	D	-	-	-	-	-	-	.521	.937	.268	.905	-	.999
	A	-	-	.928	.985	-	-	.318	.999	.602	.514	.981	.815
	CR	.992	-	-	-	-	-	.309	.997	-	.969	.814	.866

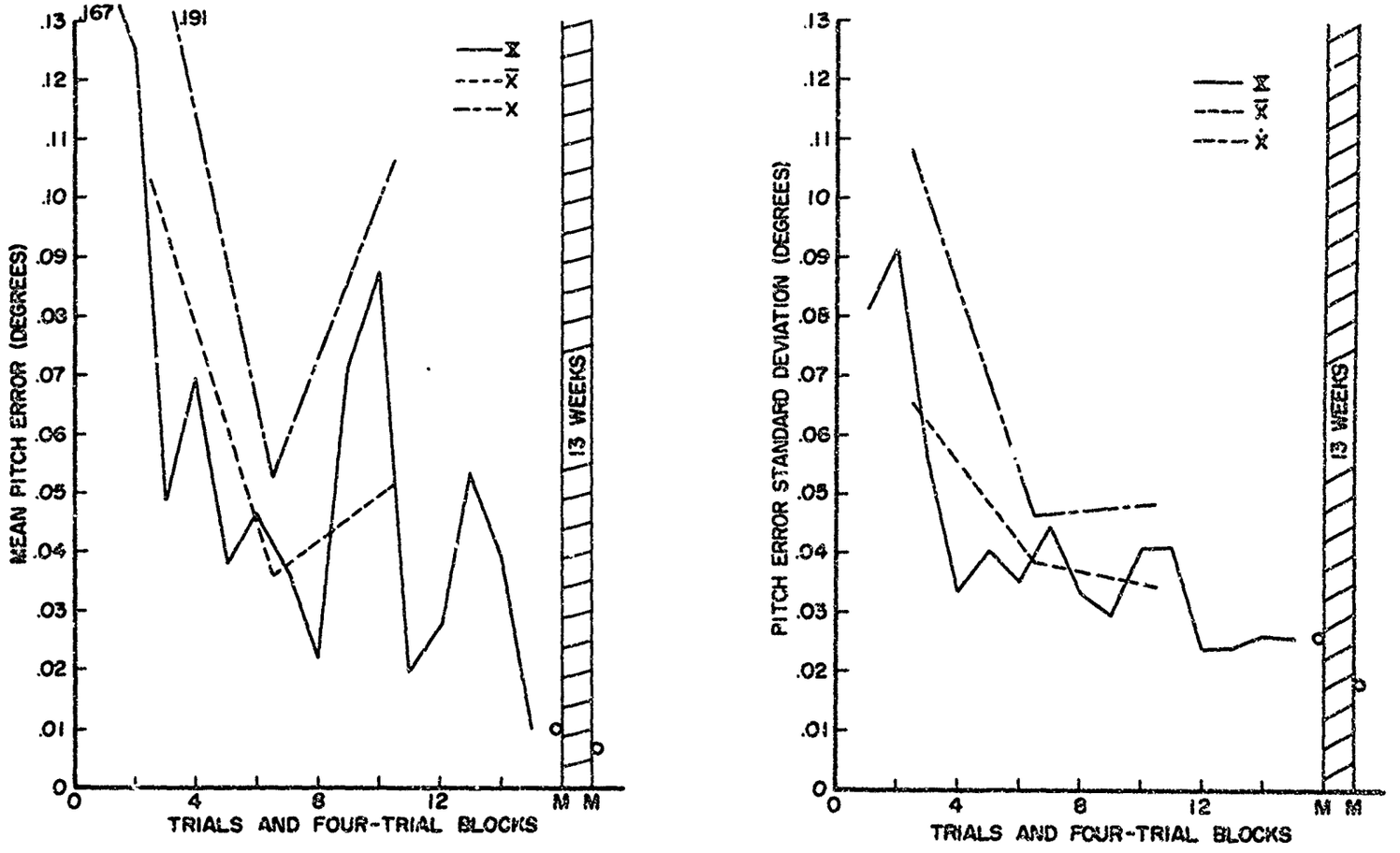
† Dash indicates value exceeding .9995
NA Necessary reference data not available

APPENDIX IV

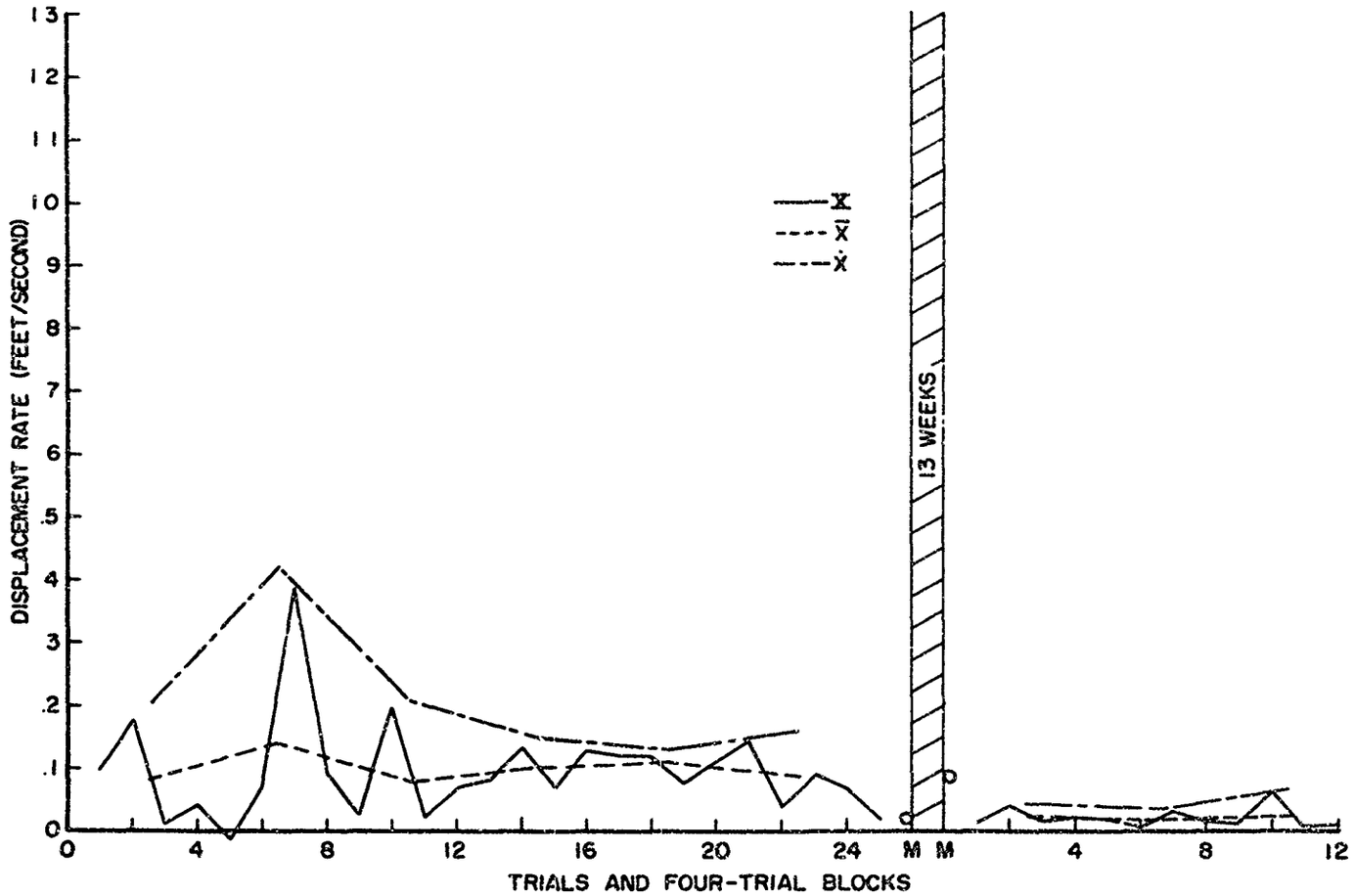
Graphs of Data from P-131



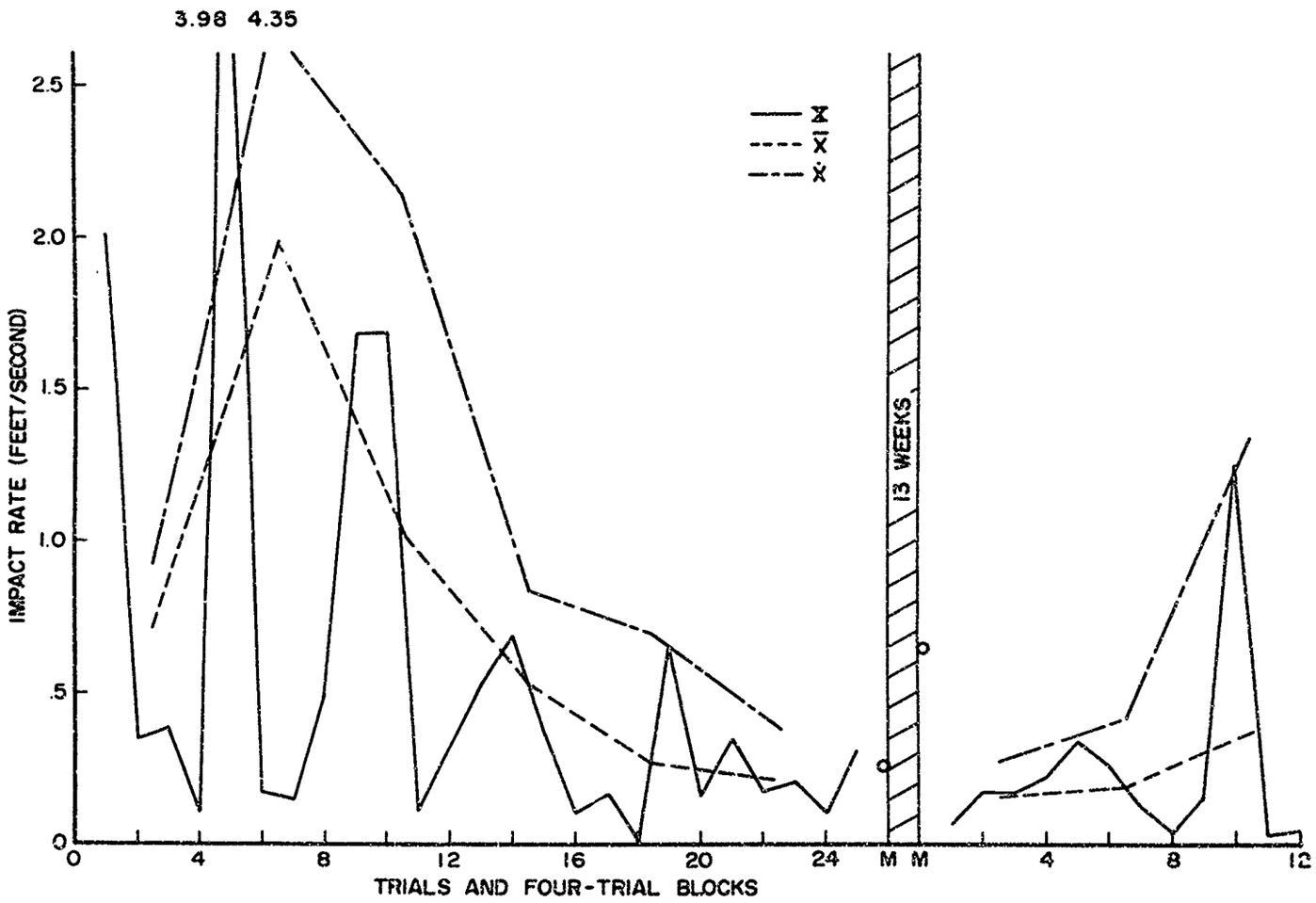
TRANSLUNAR INSERTION: VELOCITY ERROR



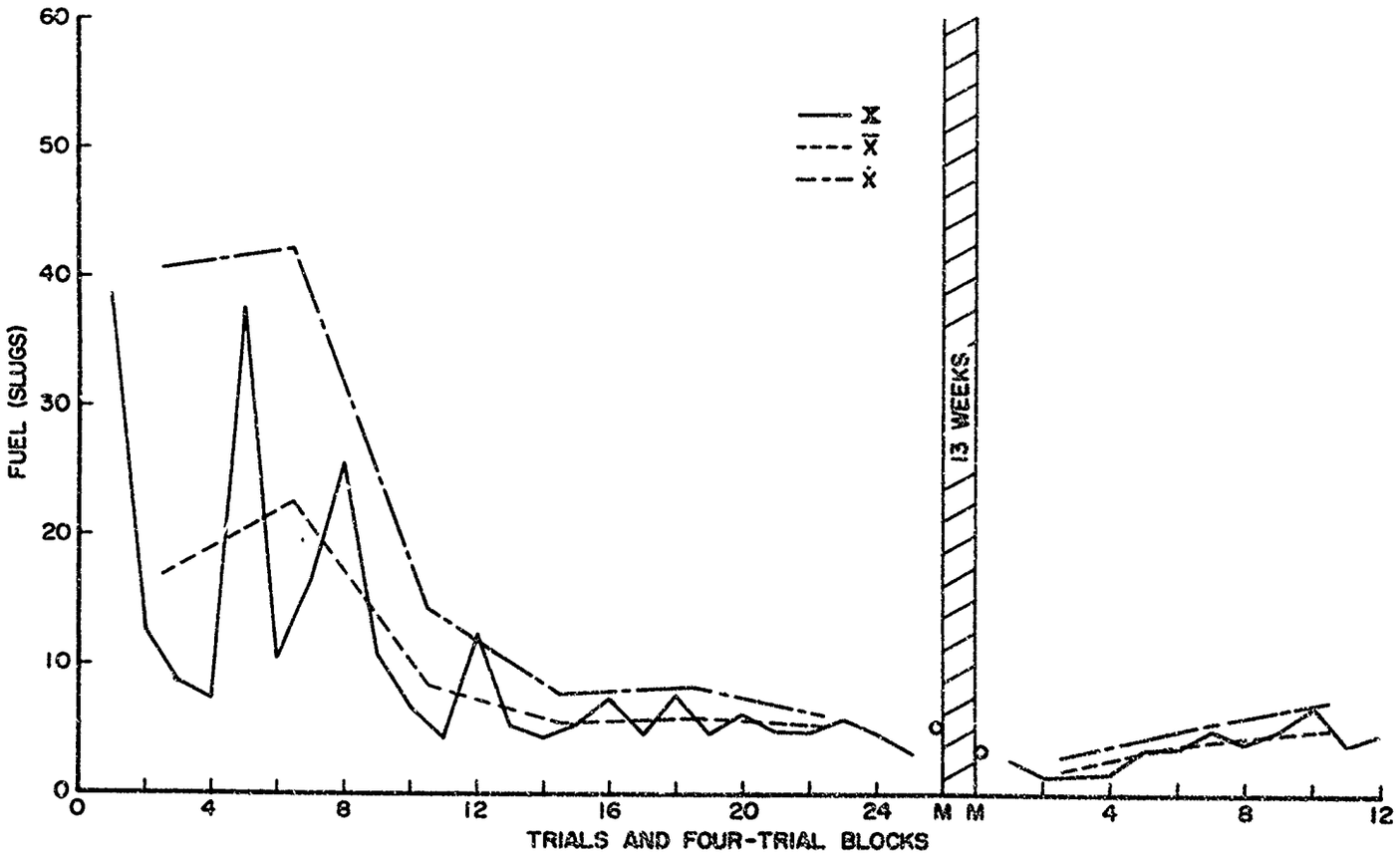
TRANSLUNAR INSERTION: PITCH ERROR



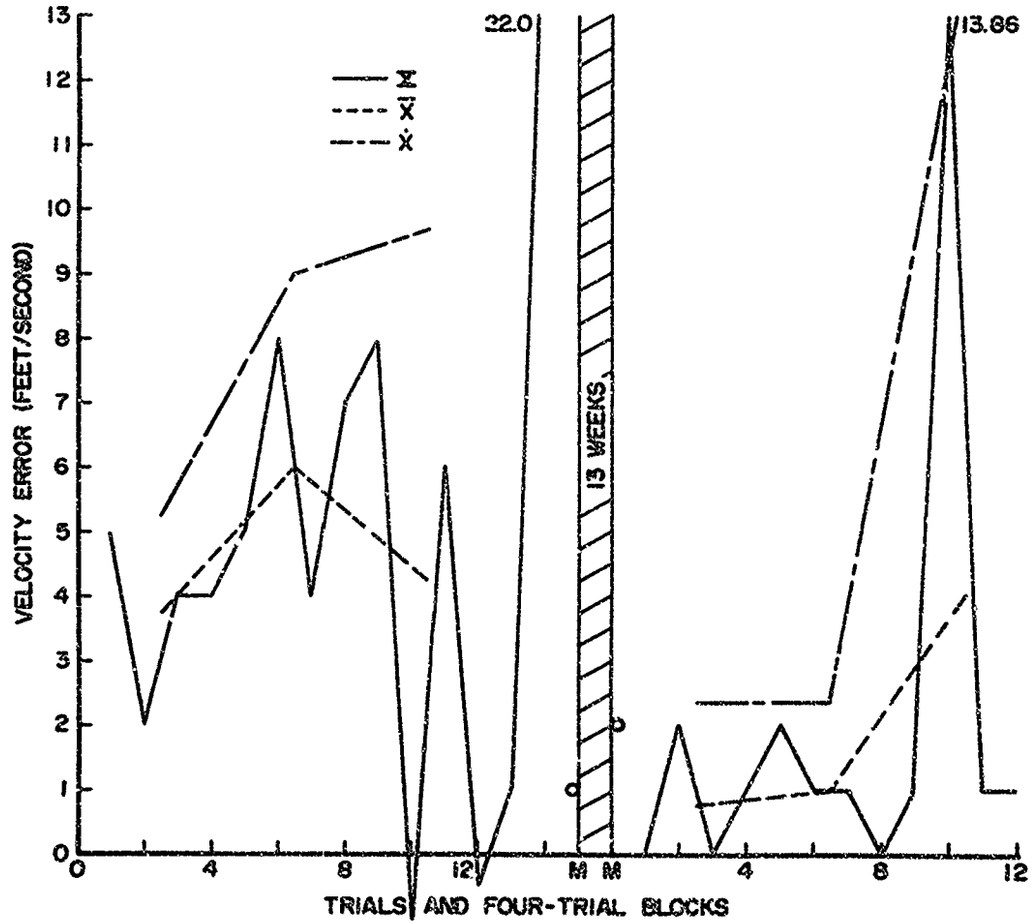
TRANSPOSITION: DISPLACEMENT RATE



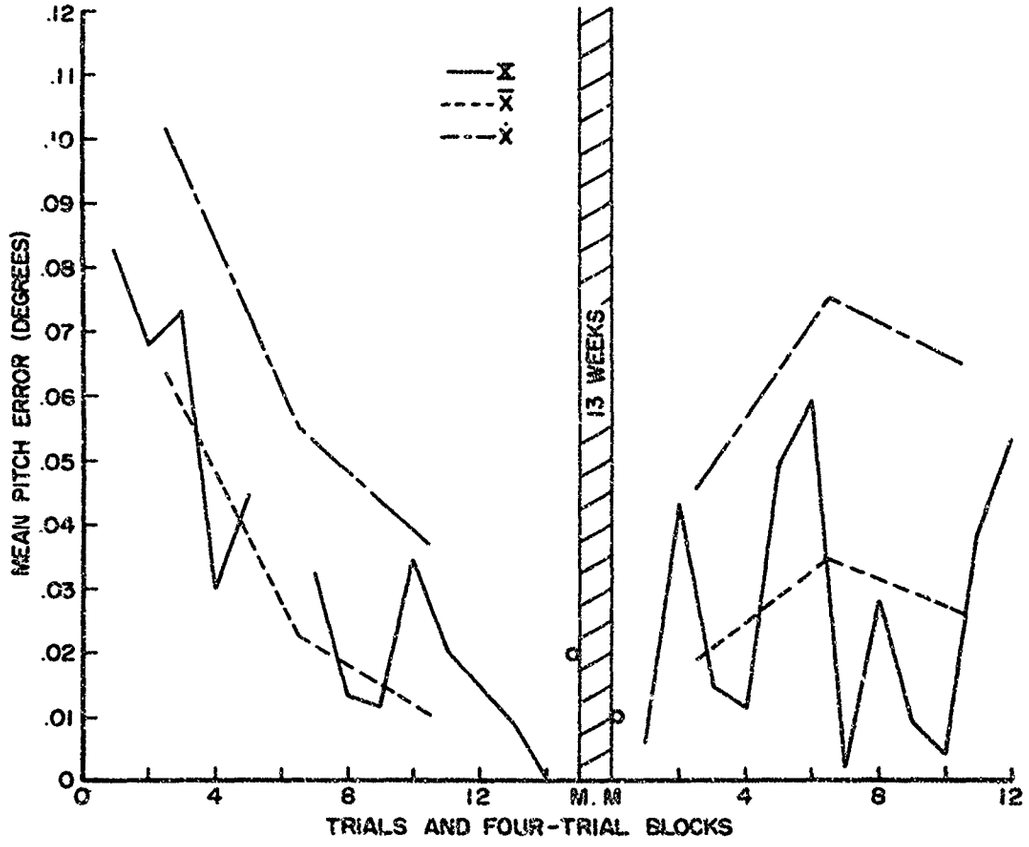
TRANSPOSITION: IMPACT RATE



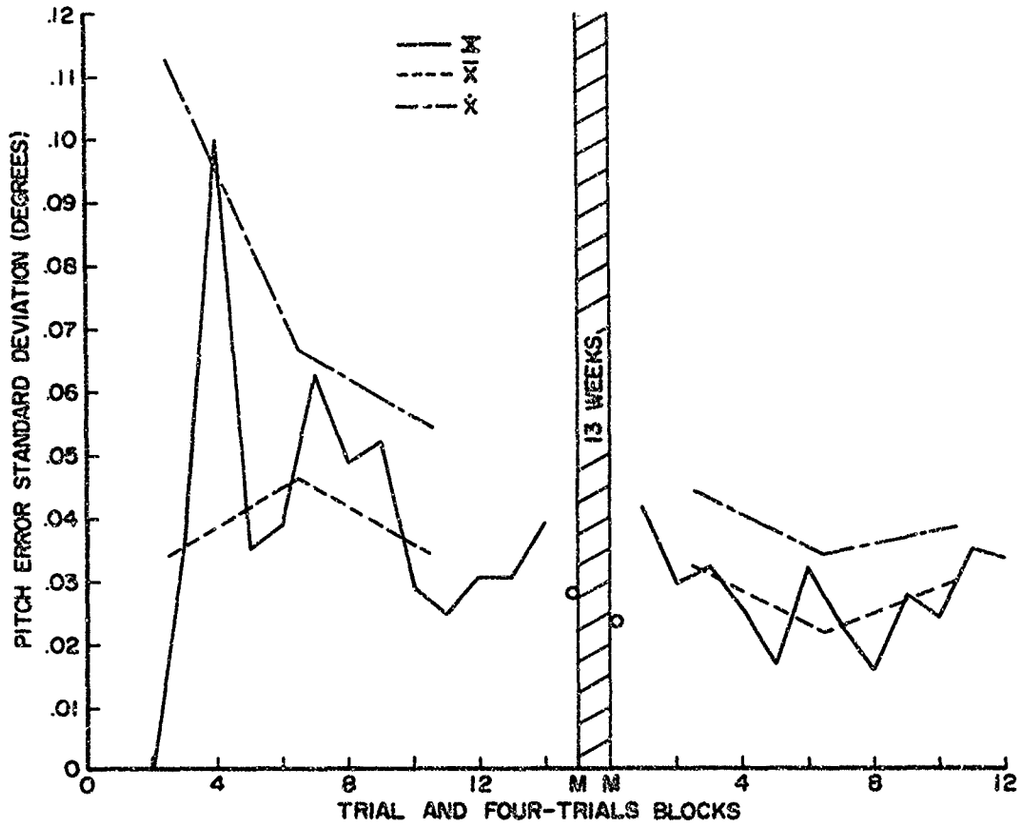
TRANSPOSITION: FUEL CONSUMED



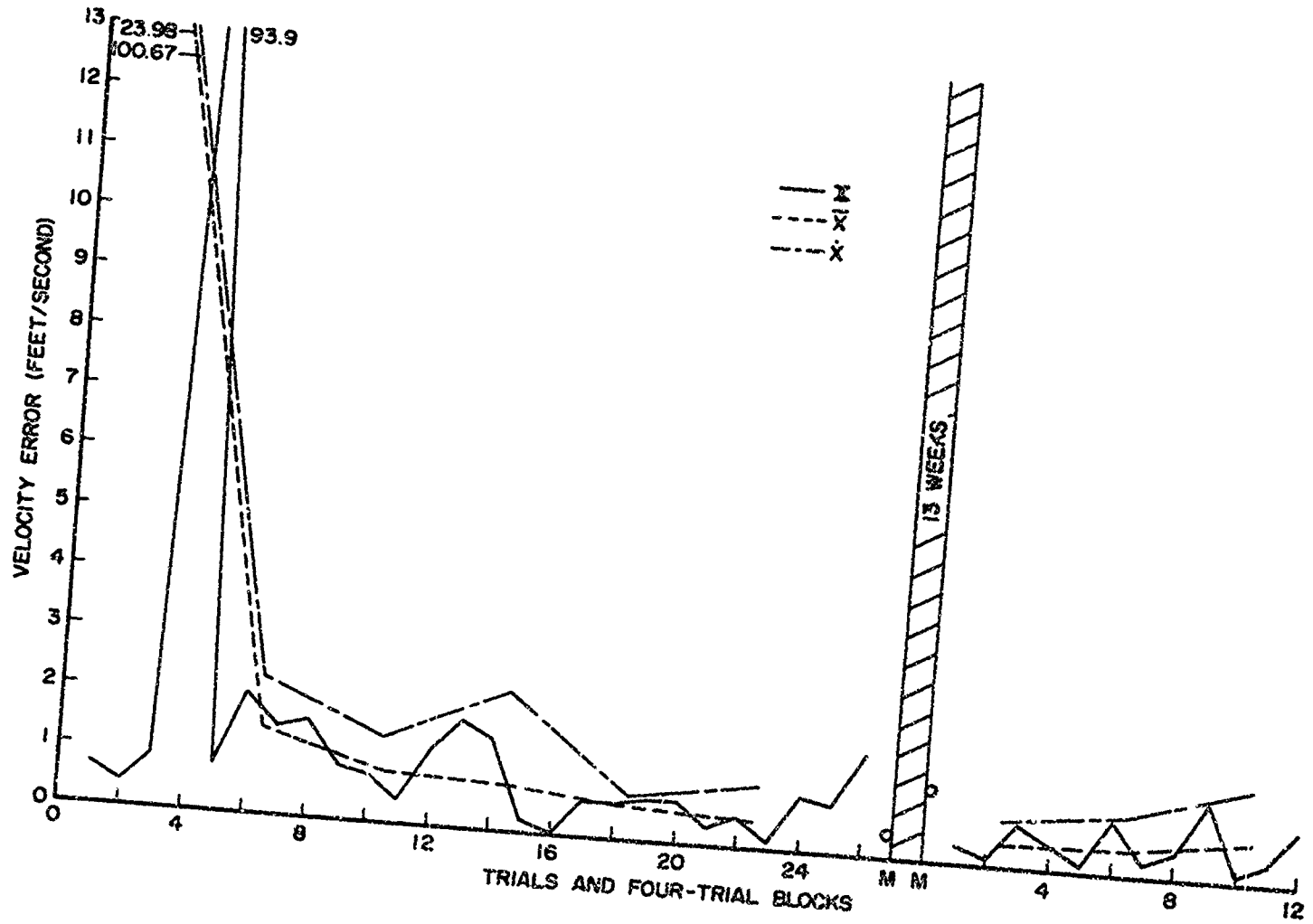
LUNAR ORBIT INSERTION: VELOCITY ERROR



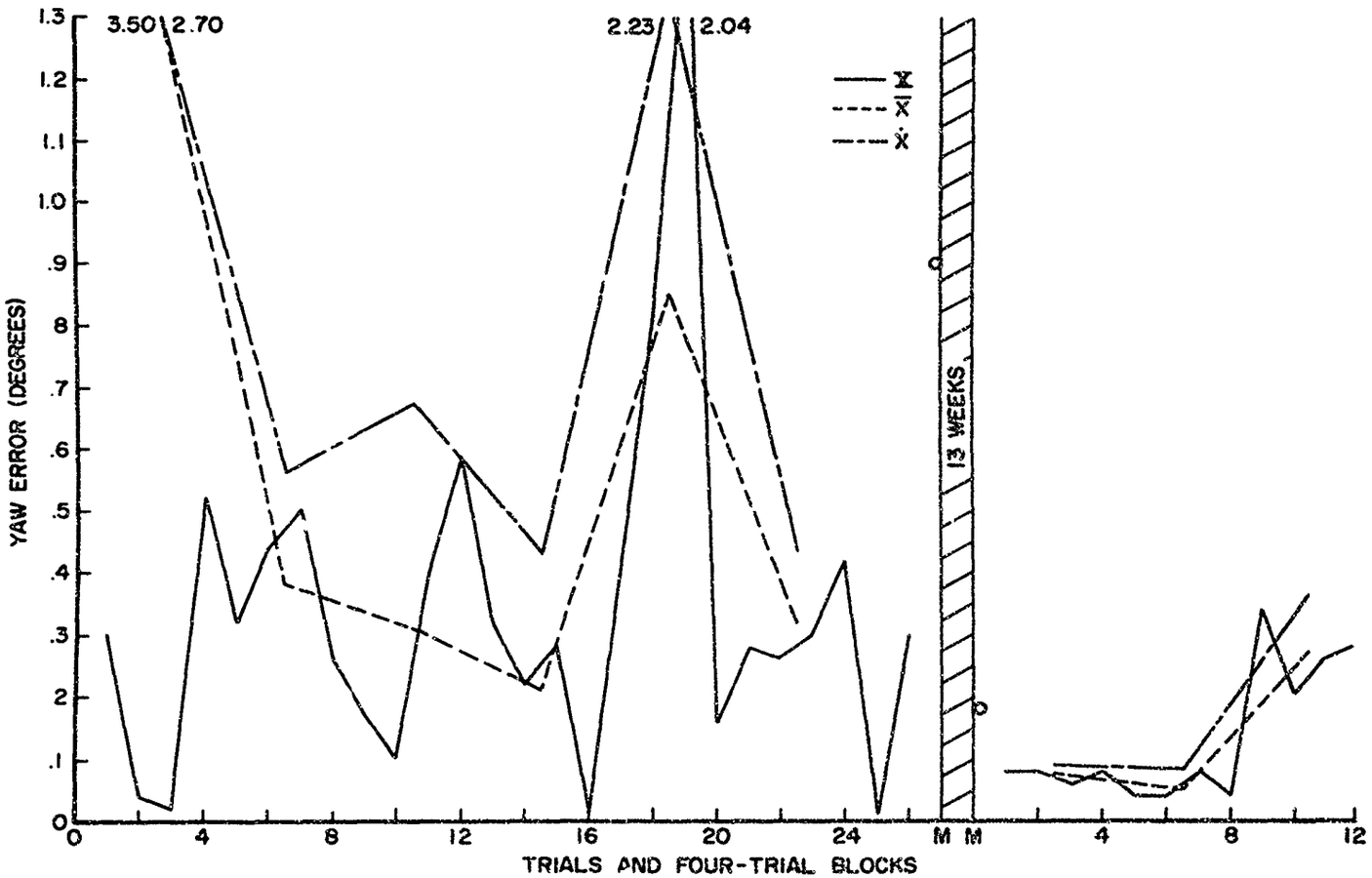
LUNAR ORBIT INSERTION: MEAN PITCH ERROR



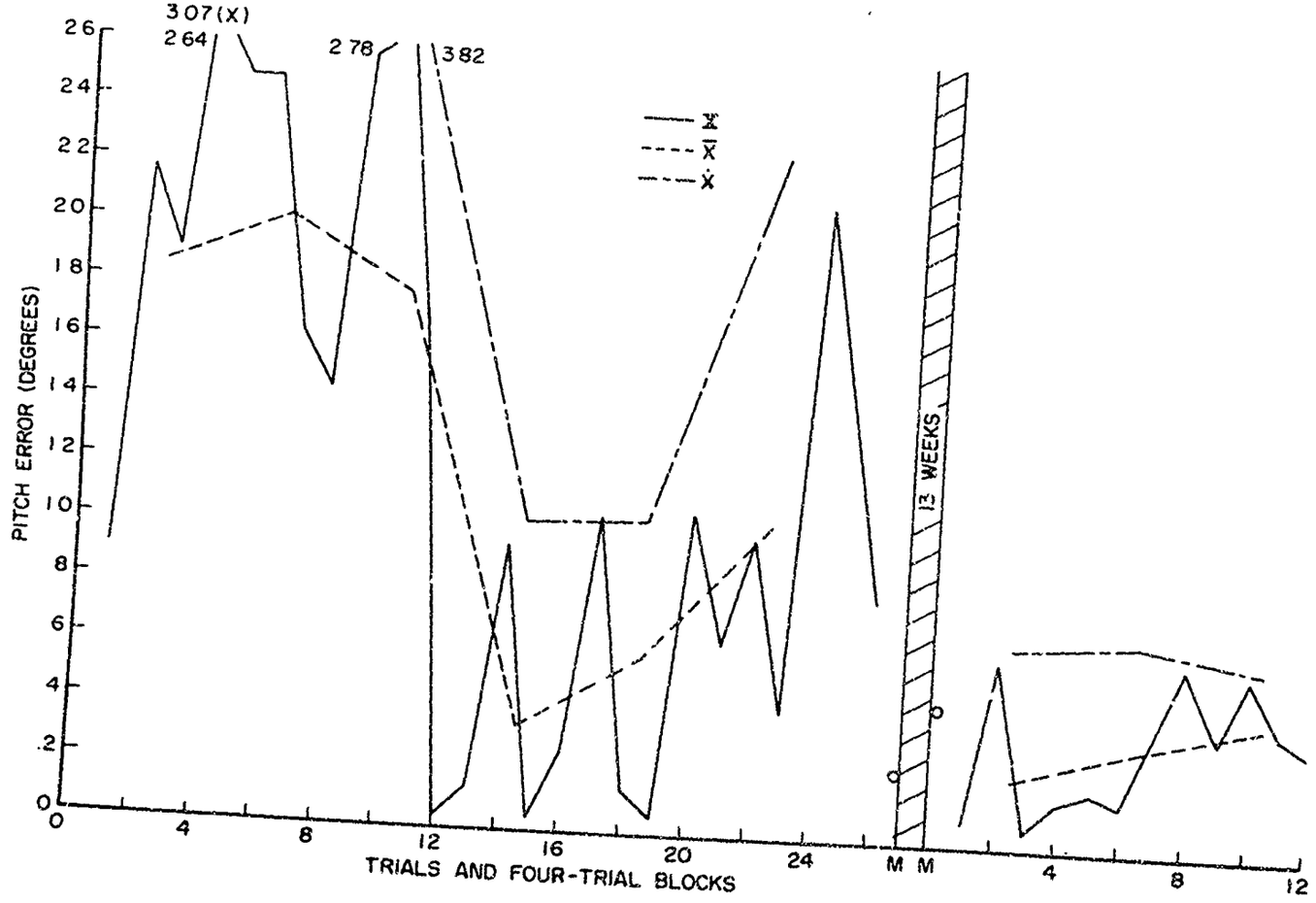
LUNAR ORBIT INSERTION: PITCH ERROR STANDARD DEVIATION



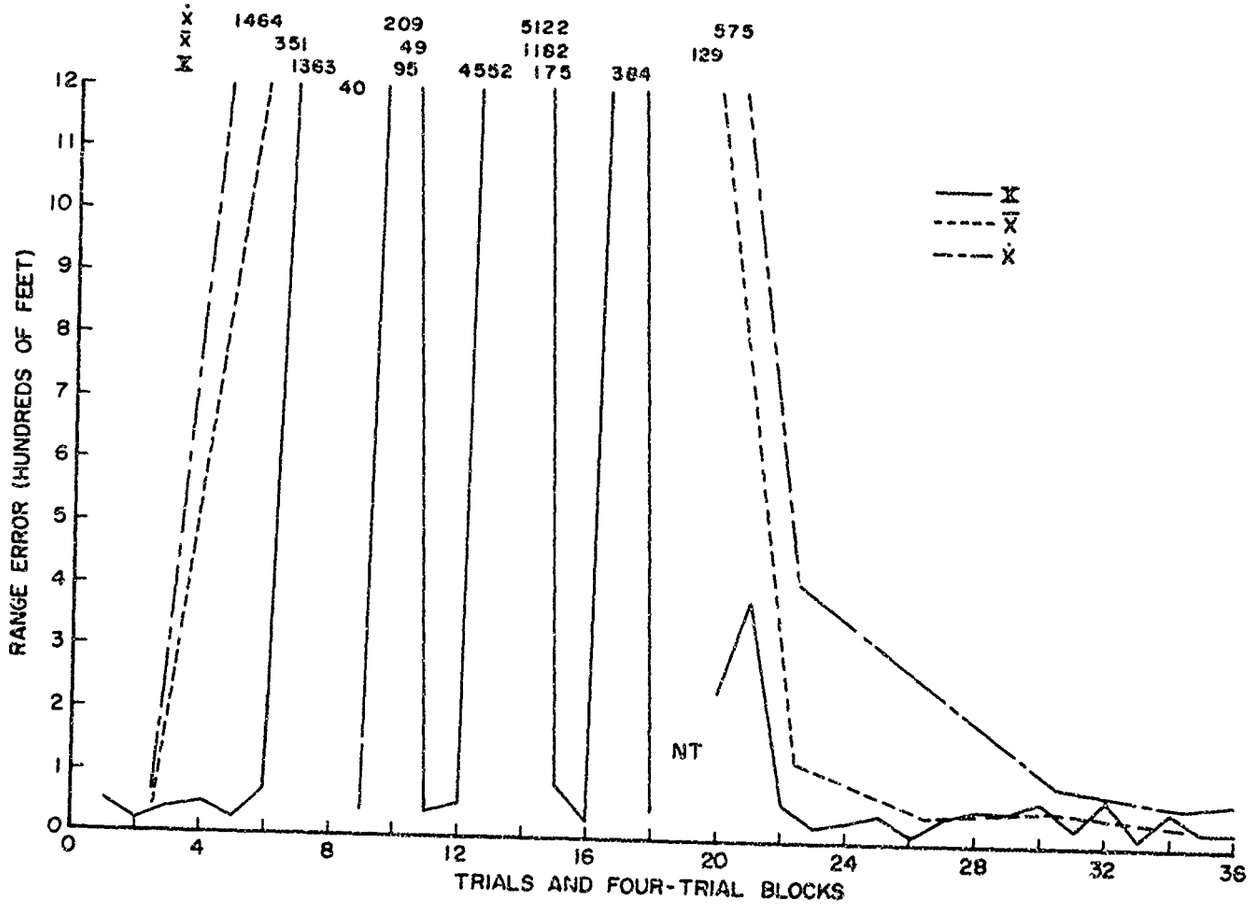
DECORBIT: VELOCITY ERROR



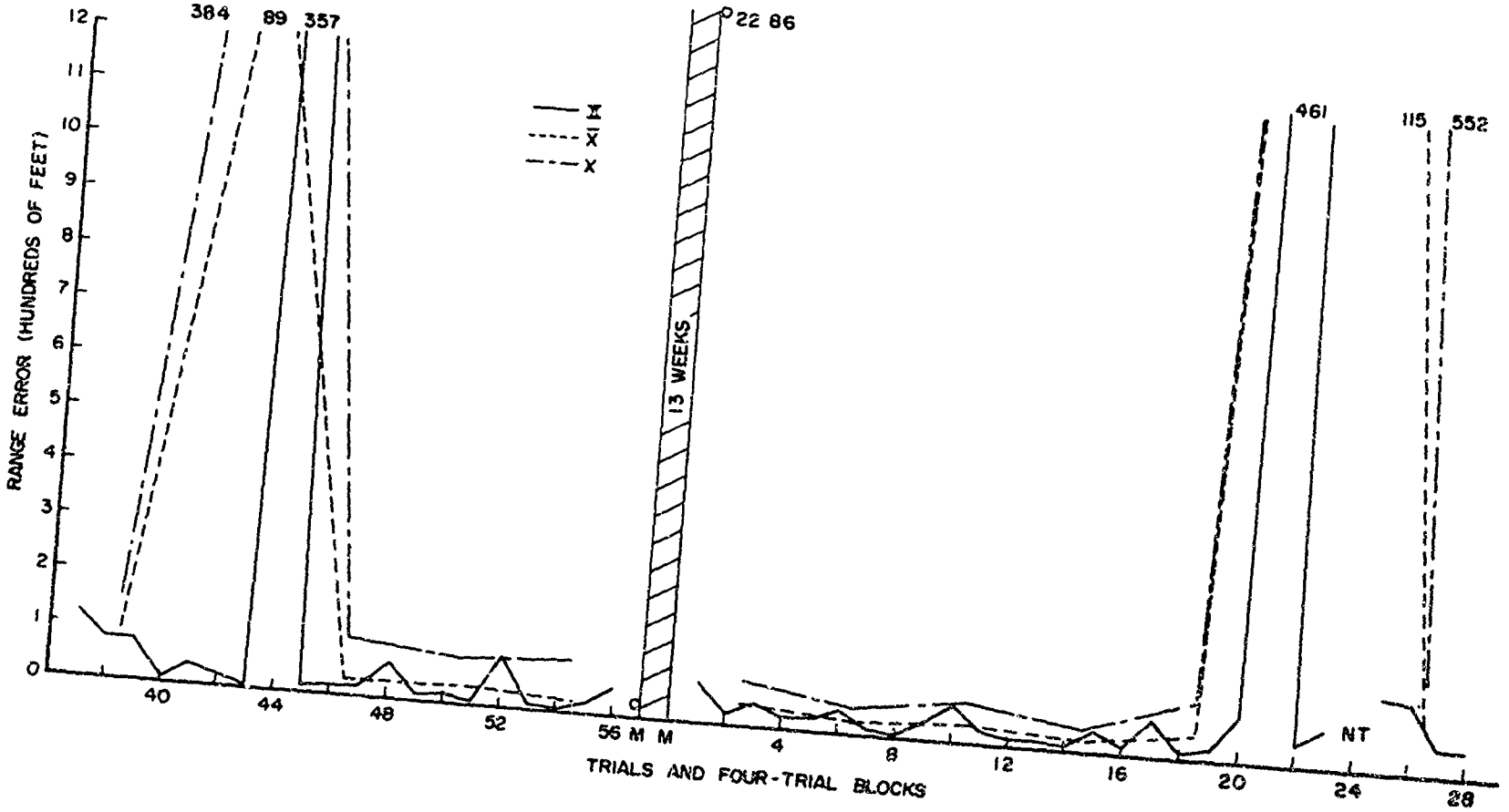
DEORBIT: YAW ERROR



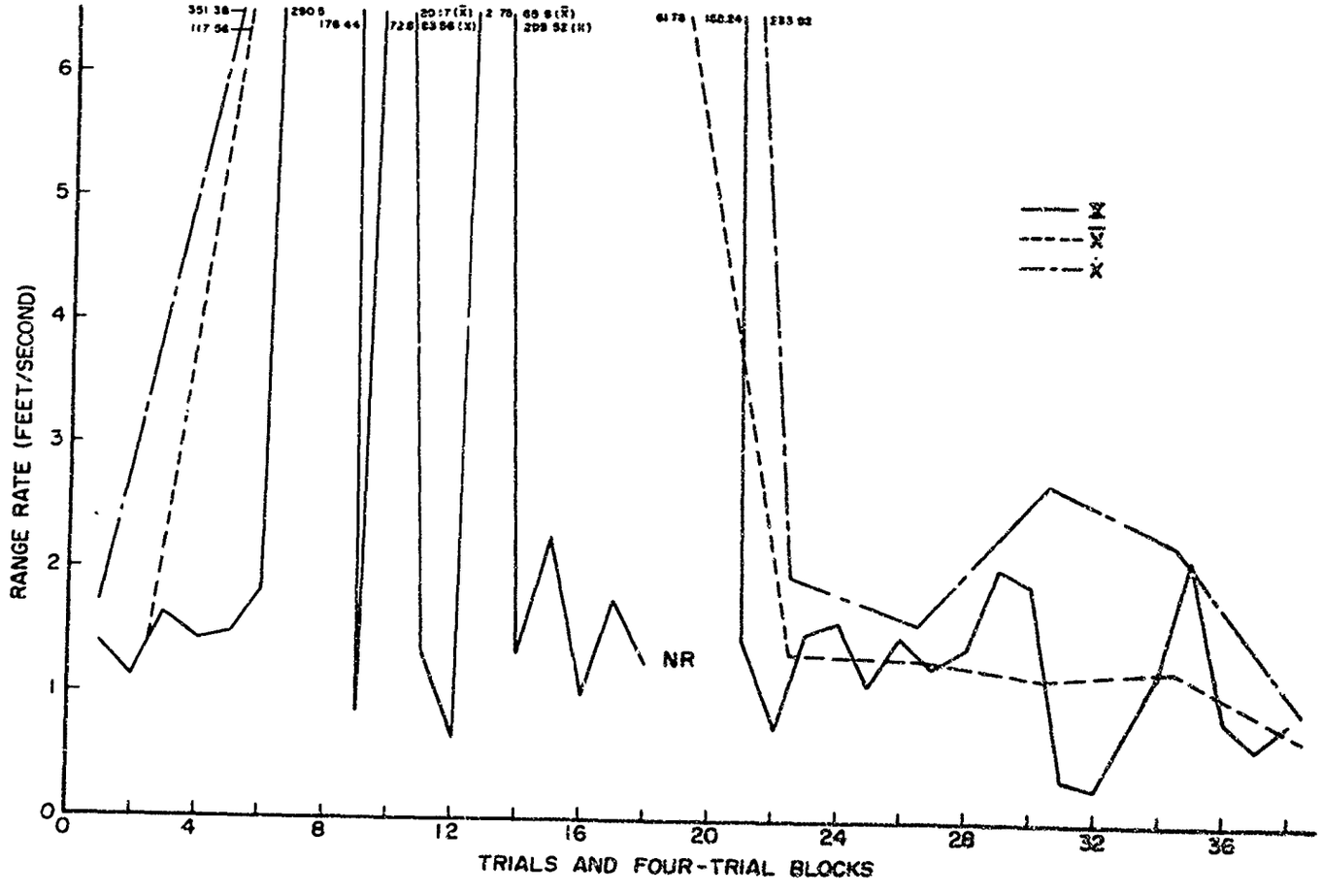
LEORRIT: PITCH ERROR



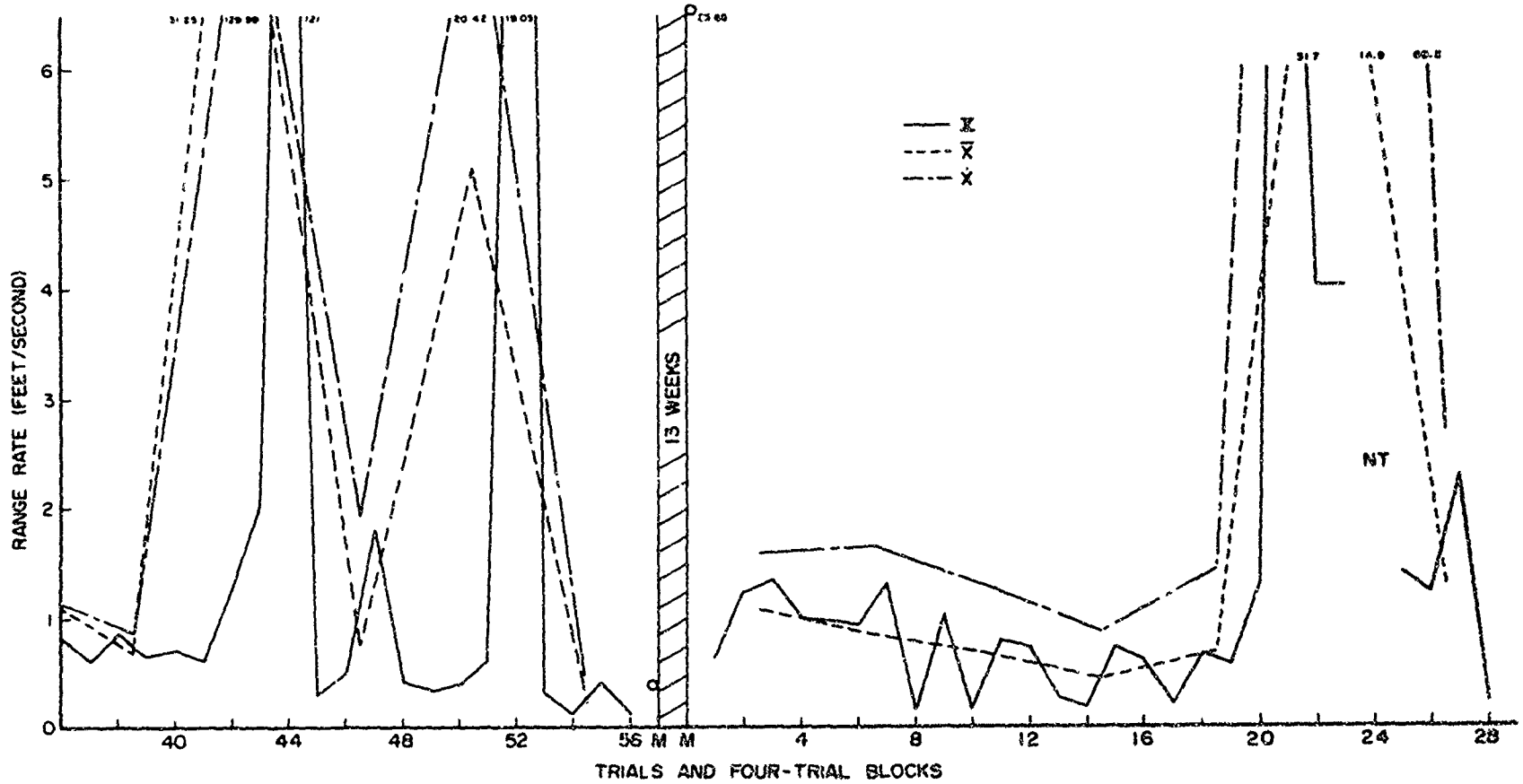
BRAKE AND HOVER: RANGE ERROR



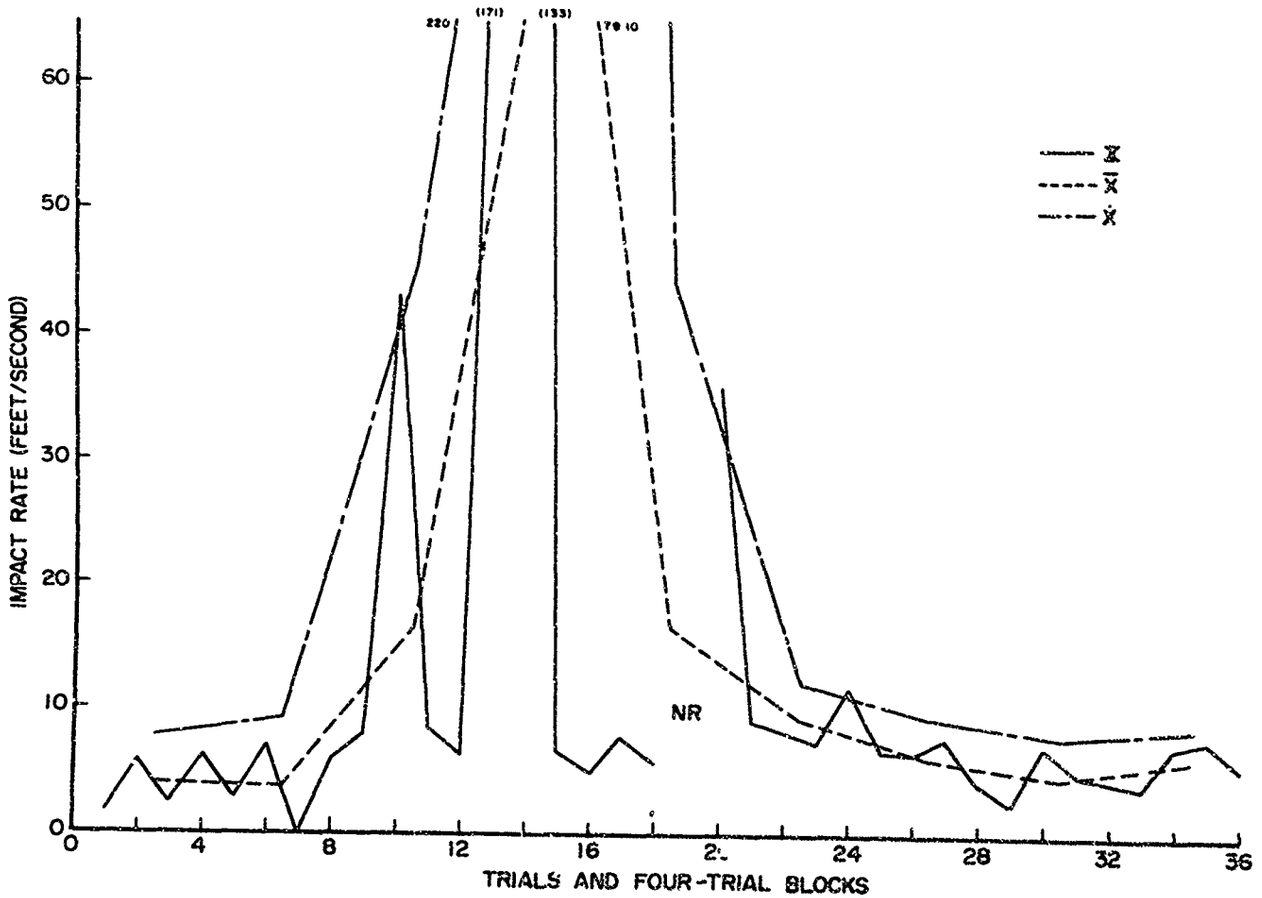
BRAKE AND HOVER: RANGE ERROR (CONTINUED)



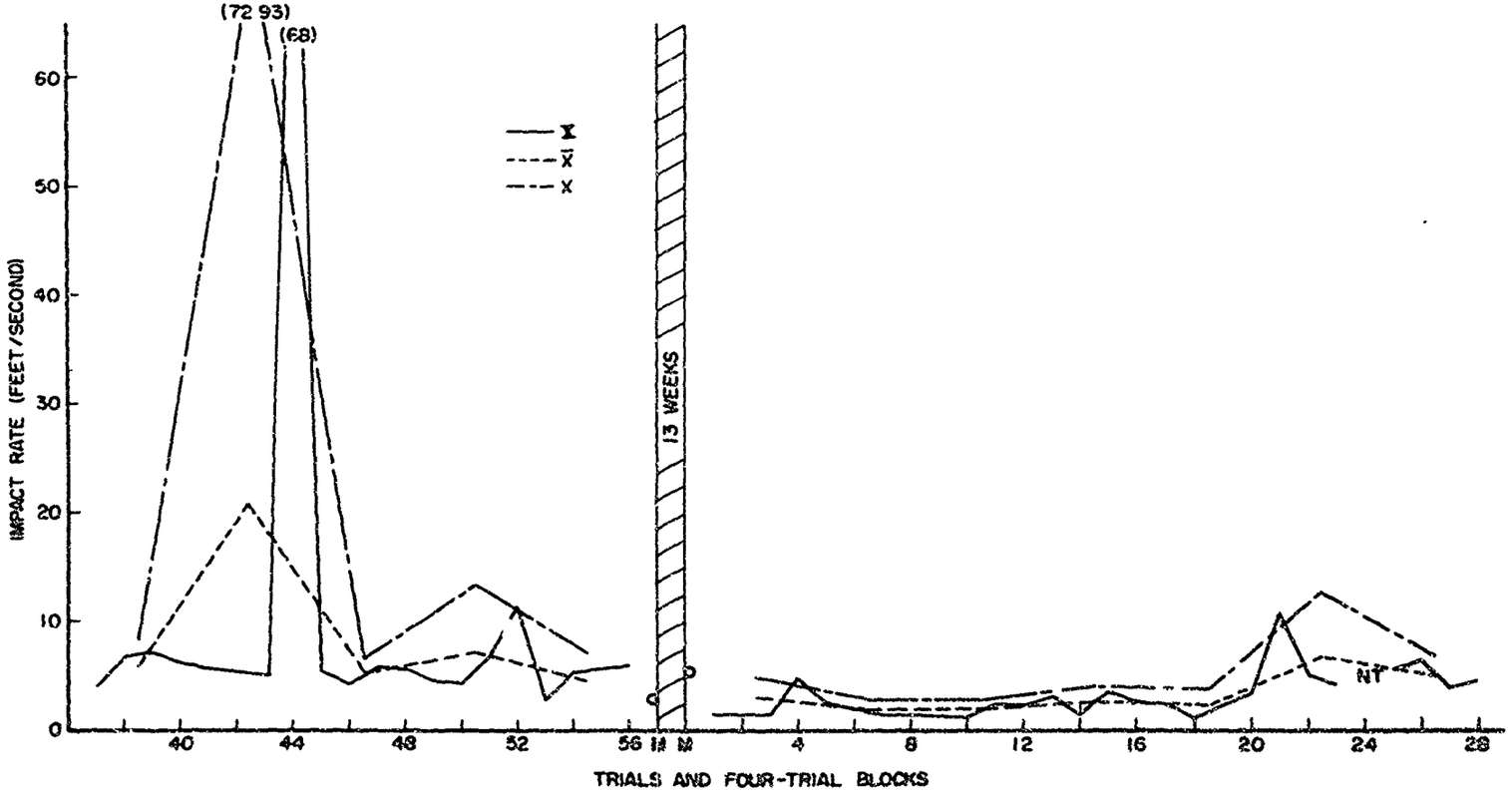
BRAKE AND HOVER: RANGE RATE



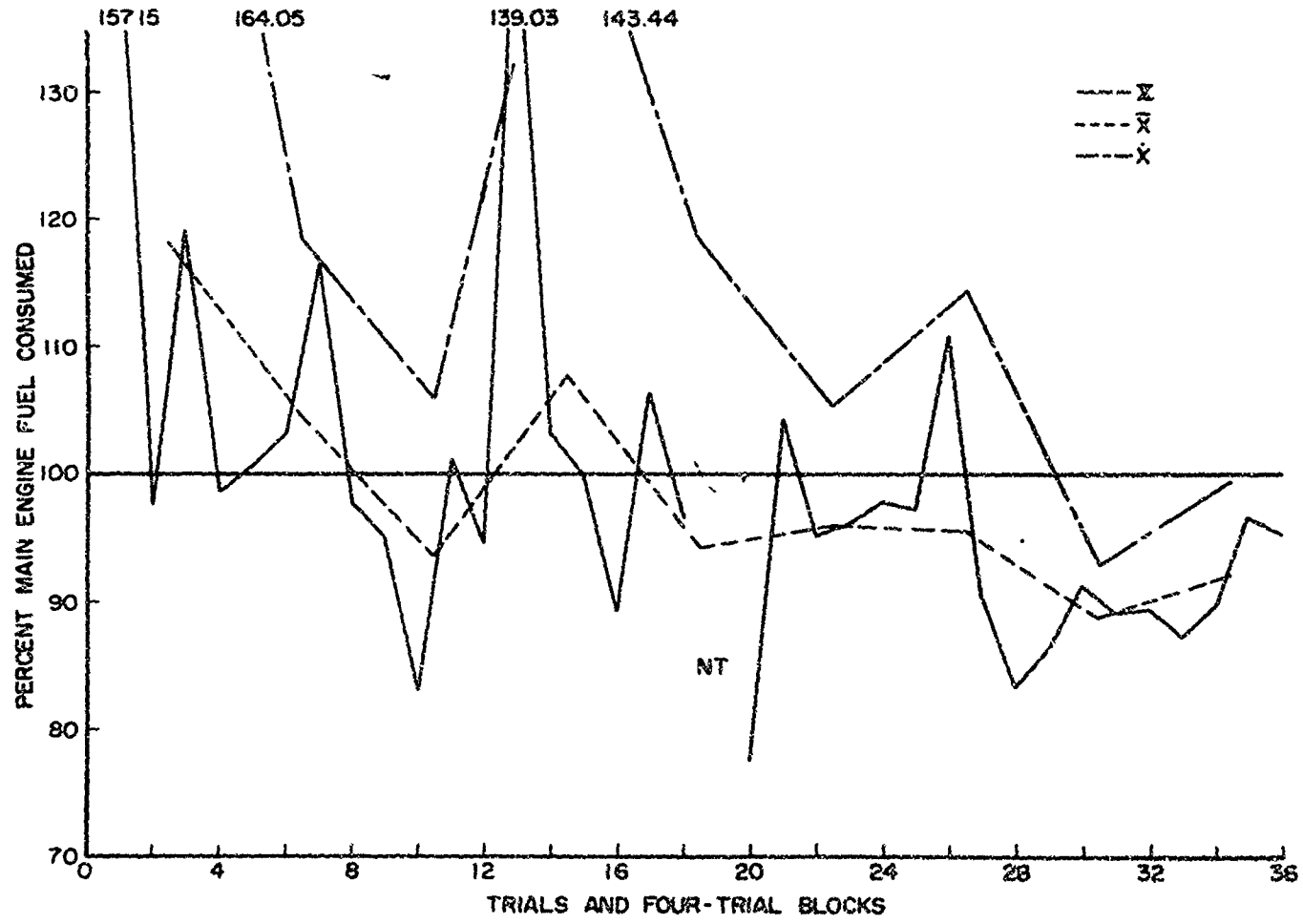
BRAKE AND HOVER: RANGE RATE (CONTINUED)



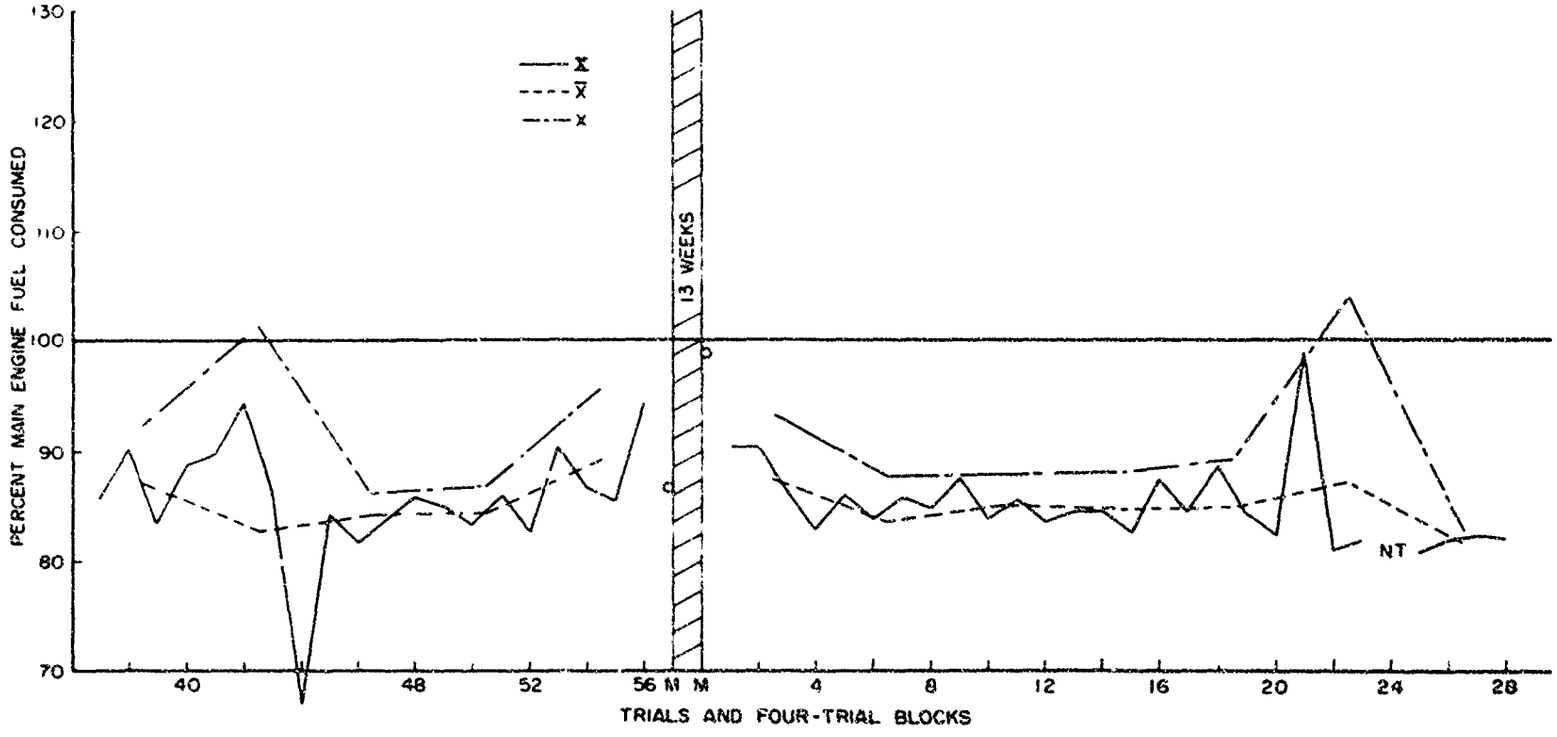
BRAKE AND HOVER: IMPACT RATE



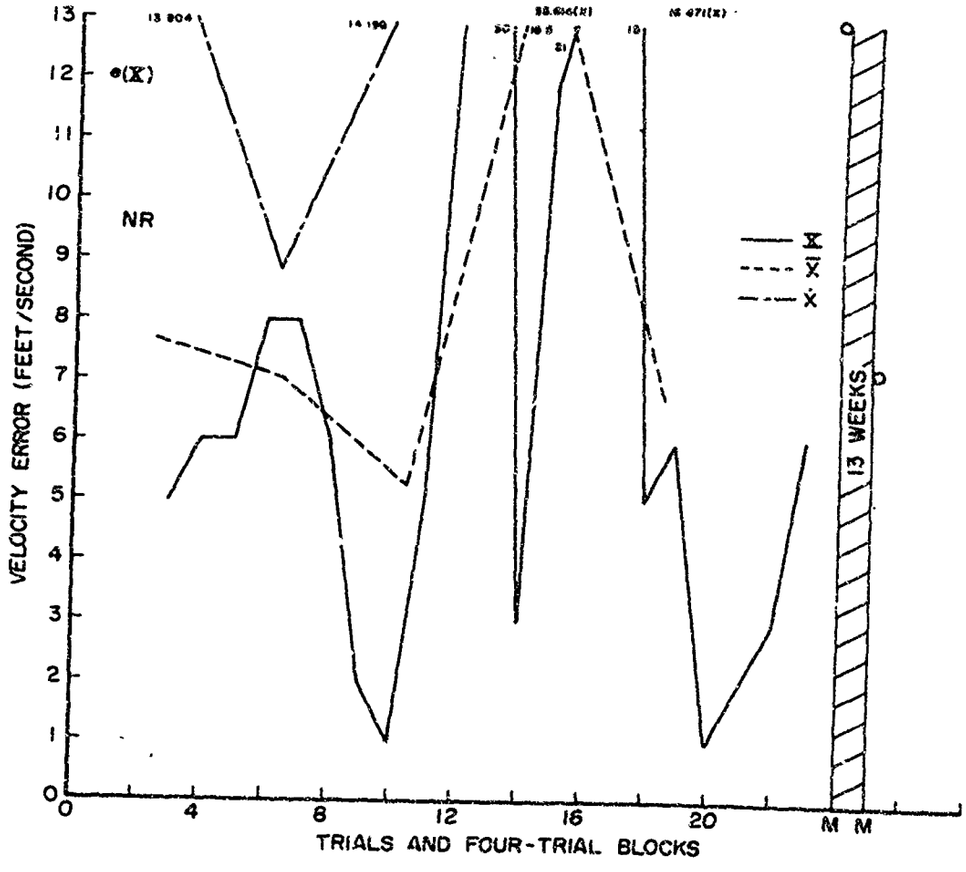
BRAKE AND HOVER: IMPACT RATE (CONTINUED)



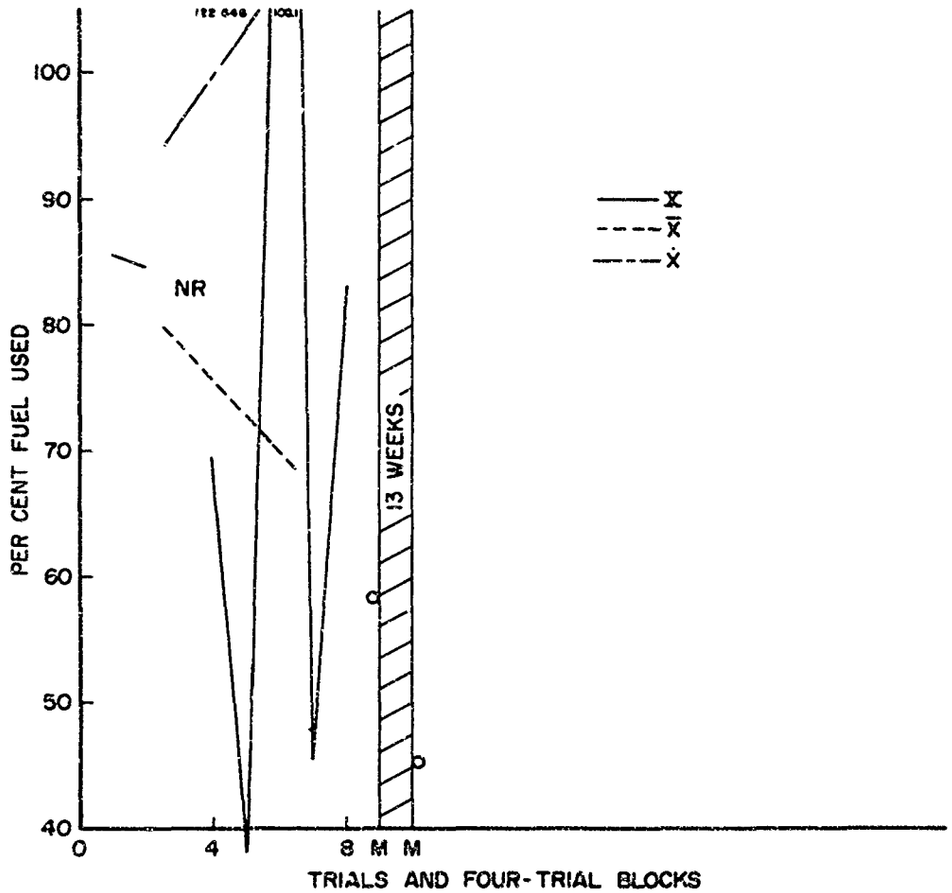
BRAKE AND HOVER: PER CENT FUEL CONSUMED



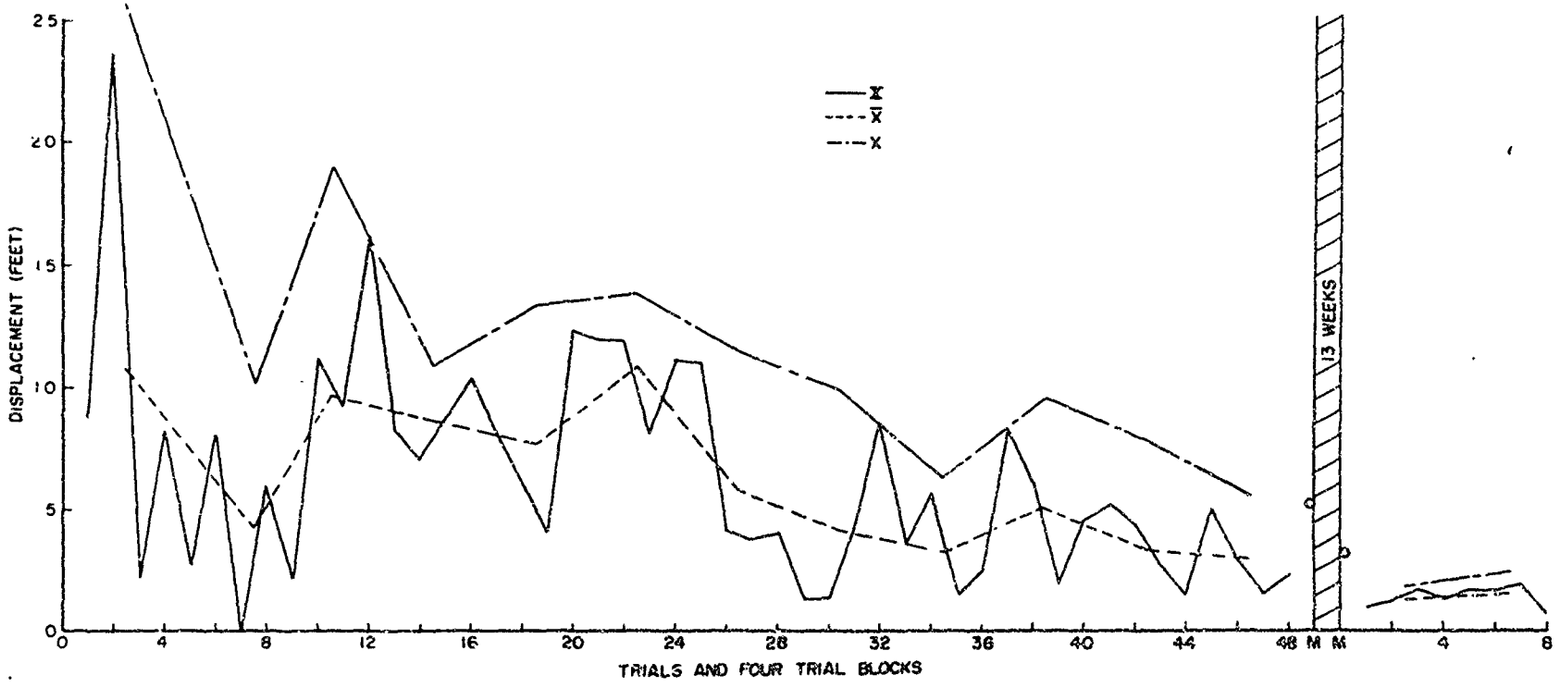
BRAKE AND HOVER: PER CENT FUEL CONSUMED (CONTINUED)



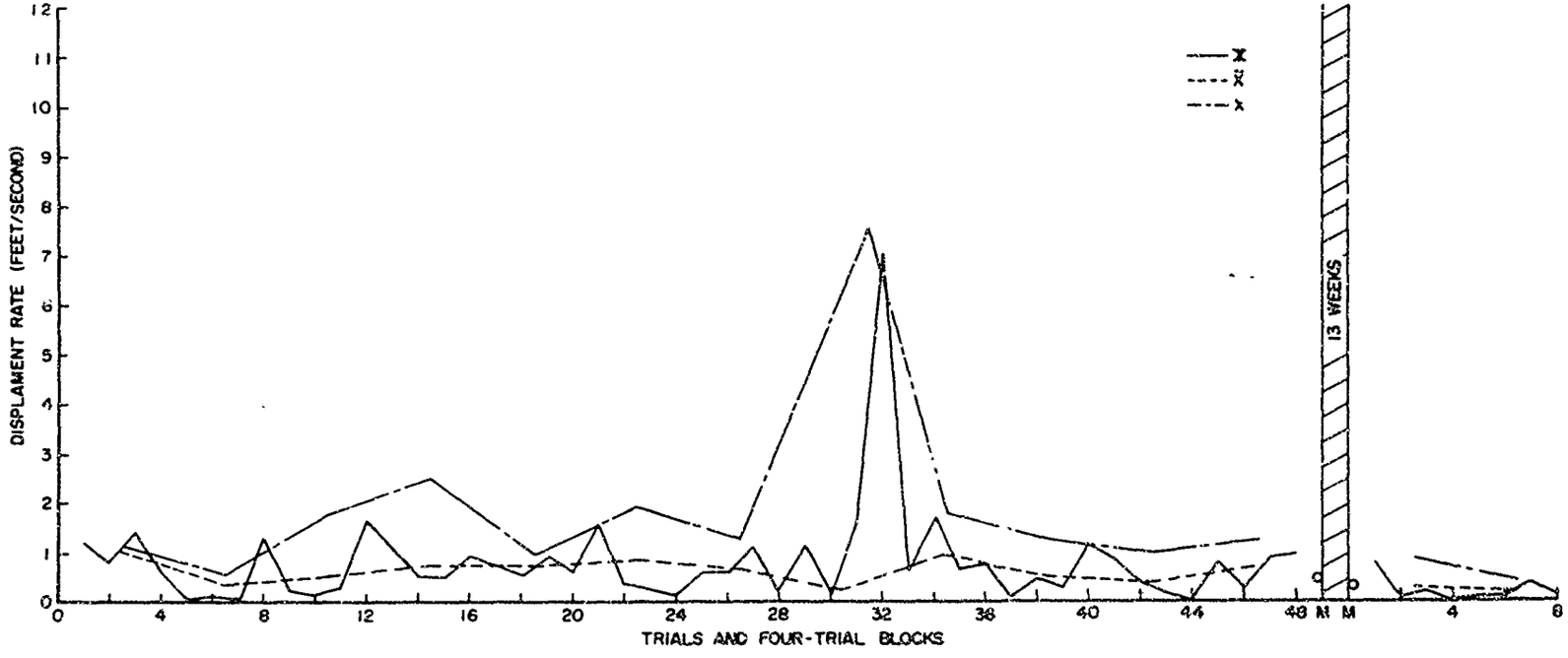
LUNAR POWERED ASCENT: VELOCITY ERROR



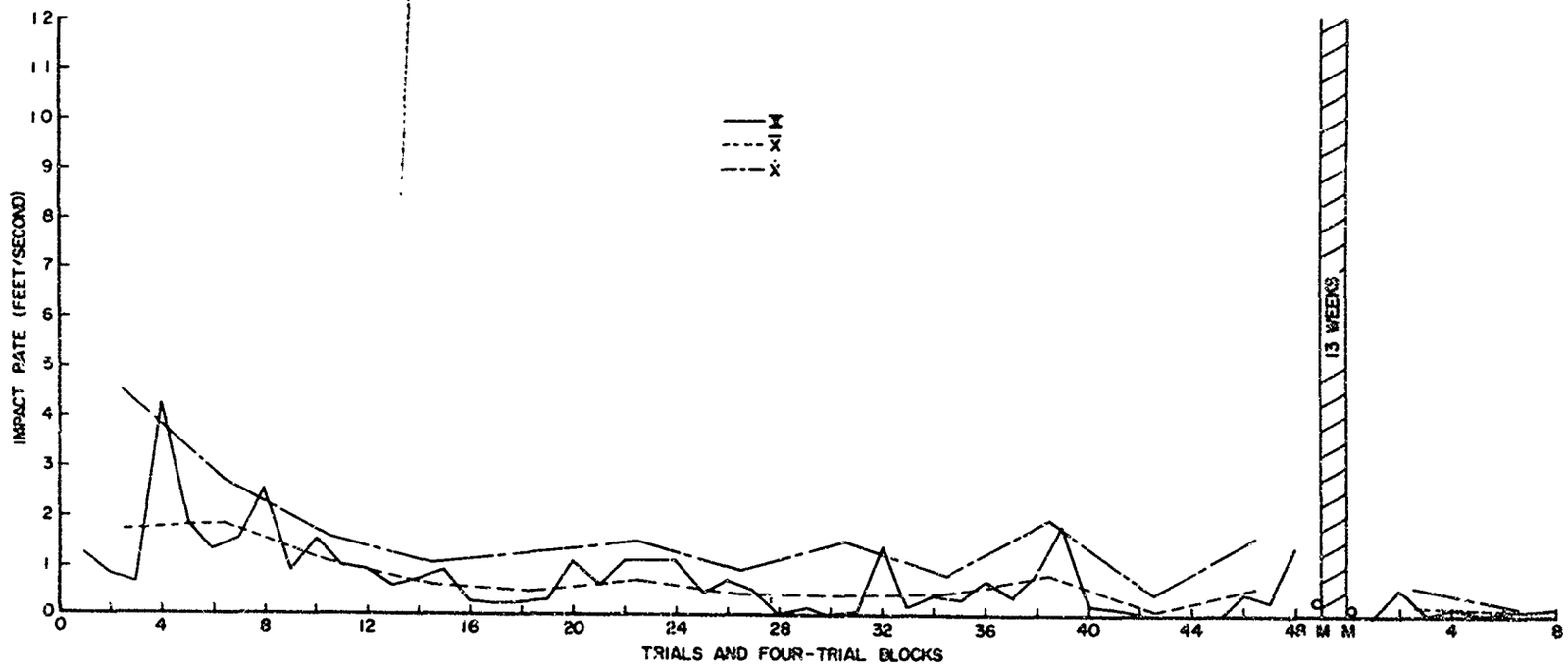
RENDEZVOUS: PER CENT FUEL CONSUMED



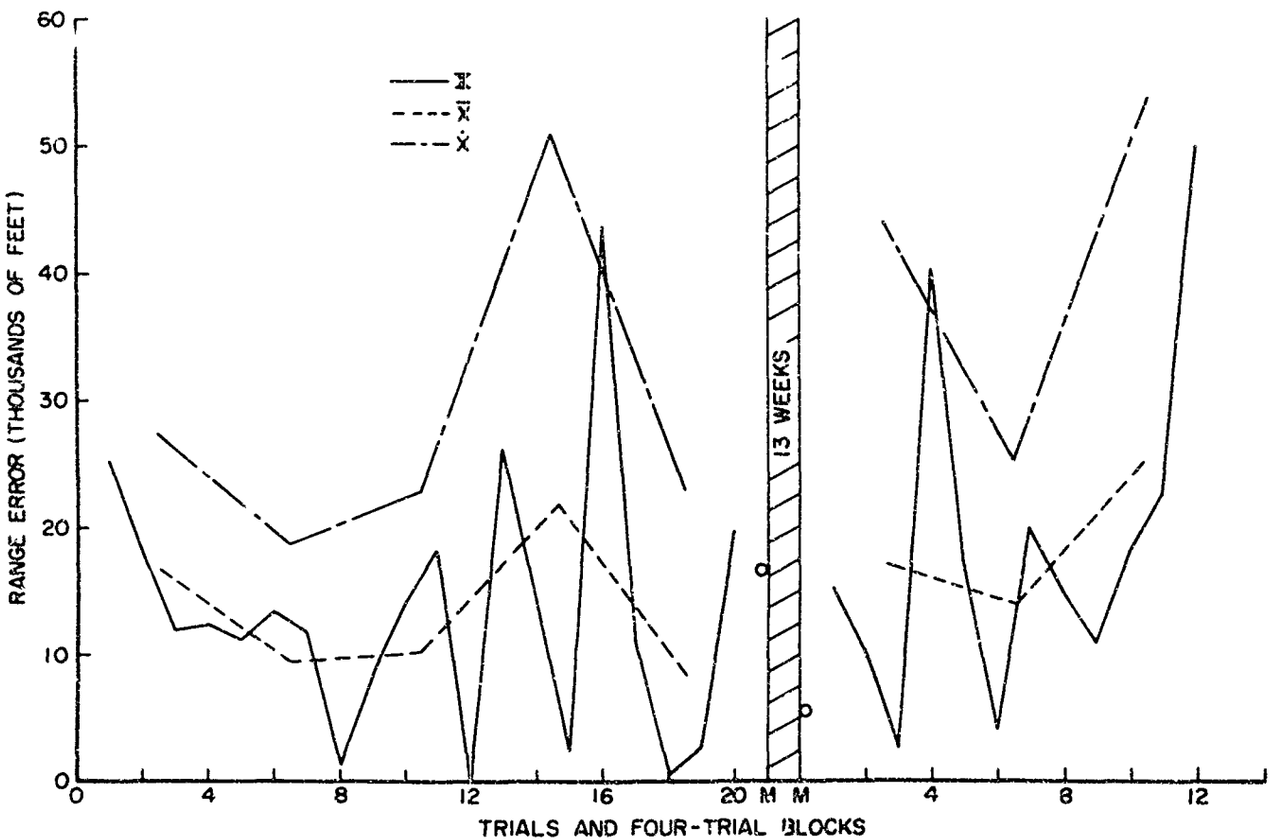
DOCKING: DISPLACEMENT



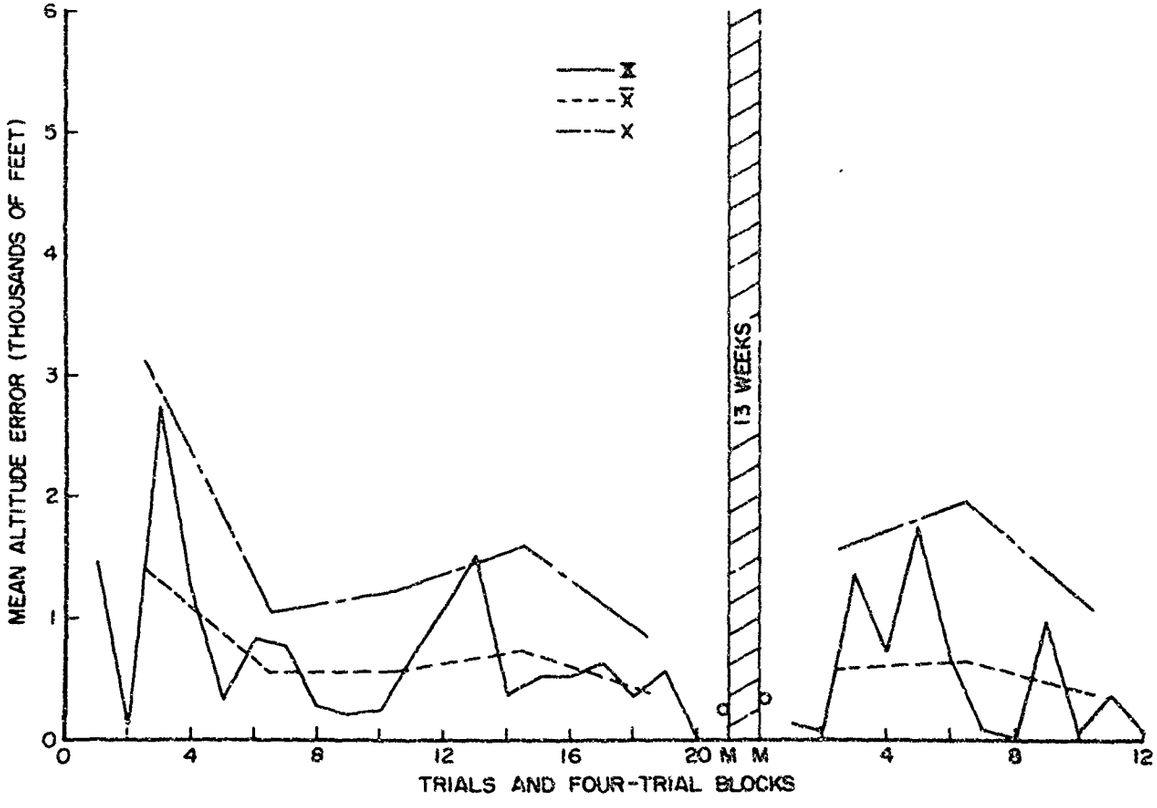
DOCKING: DISPLACEMENT RATE



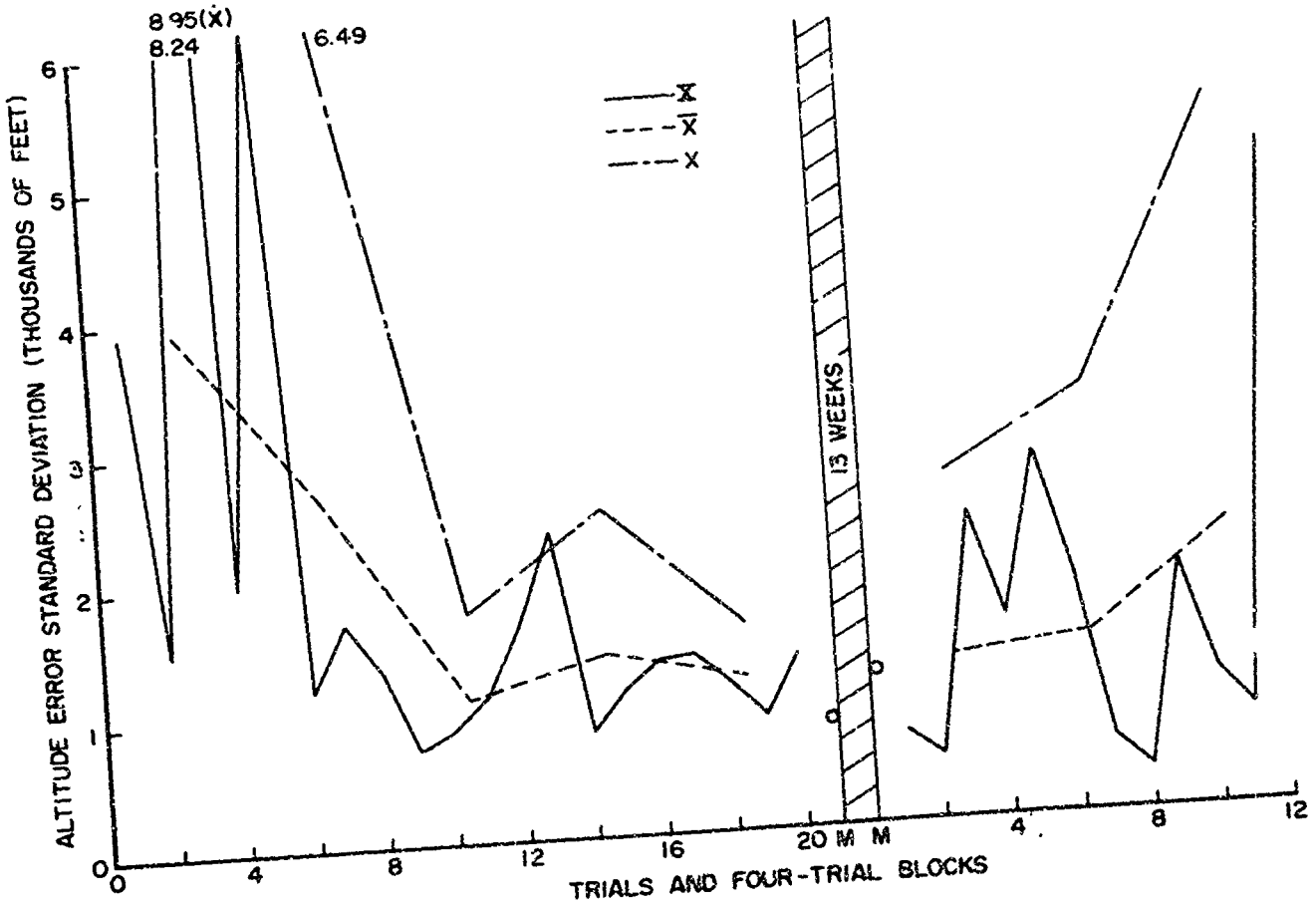
DOCKING: IMPACT RATE



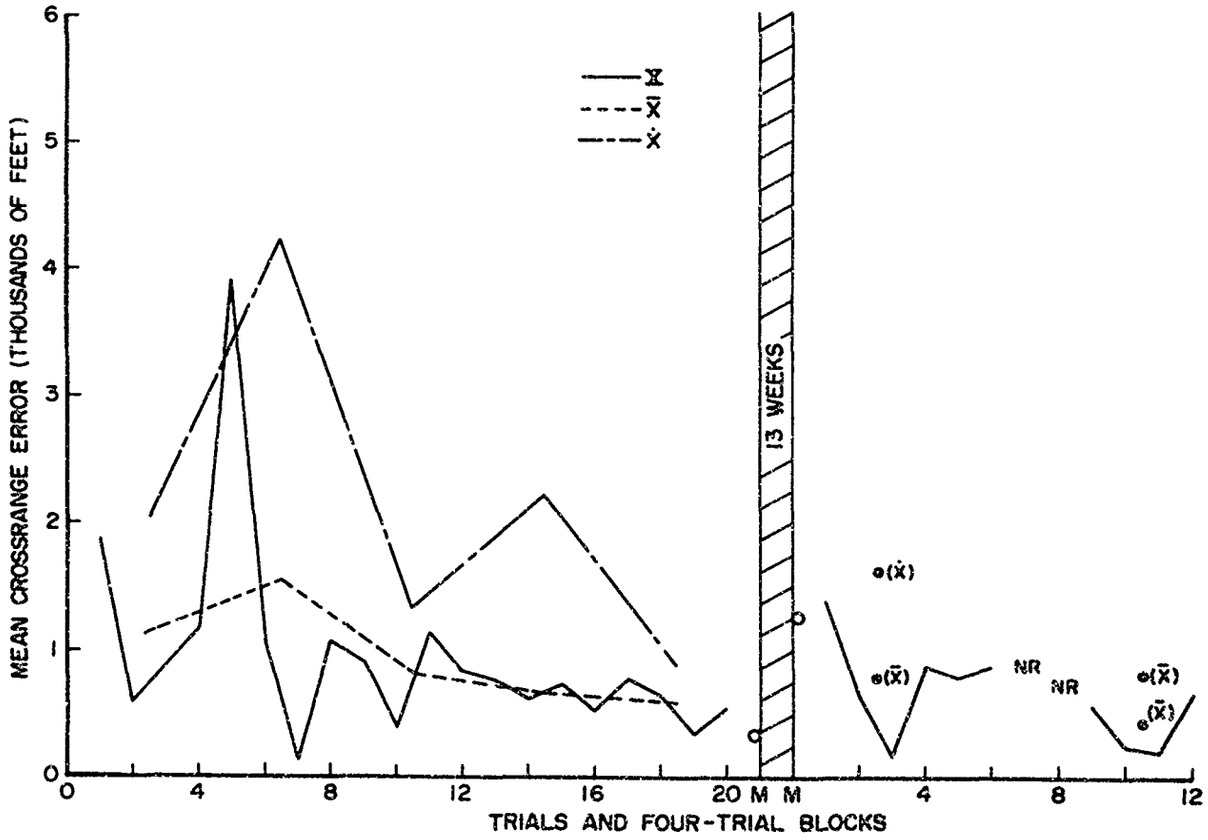
EARTH ENTRY: RANGE ERROR



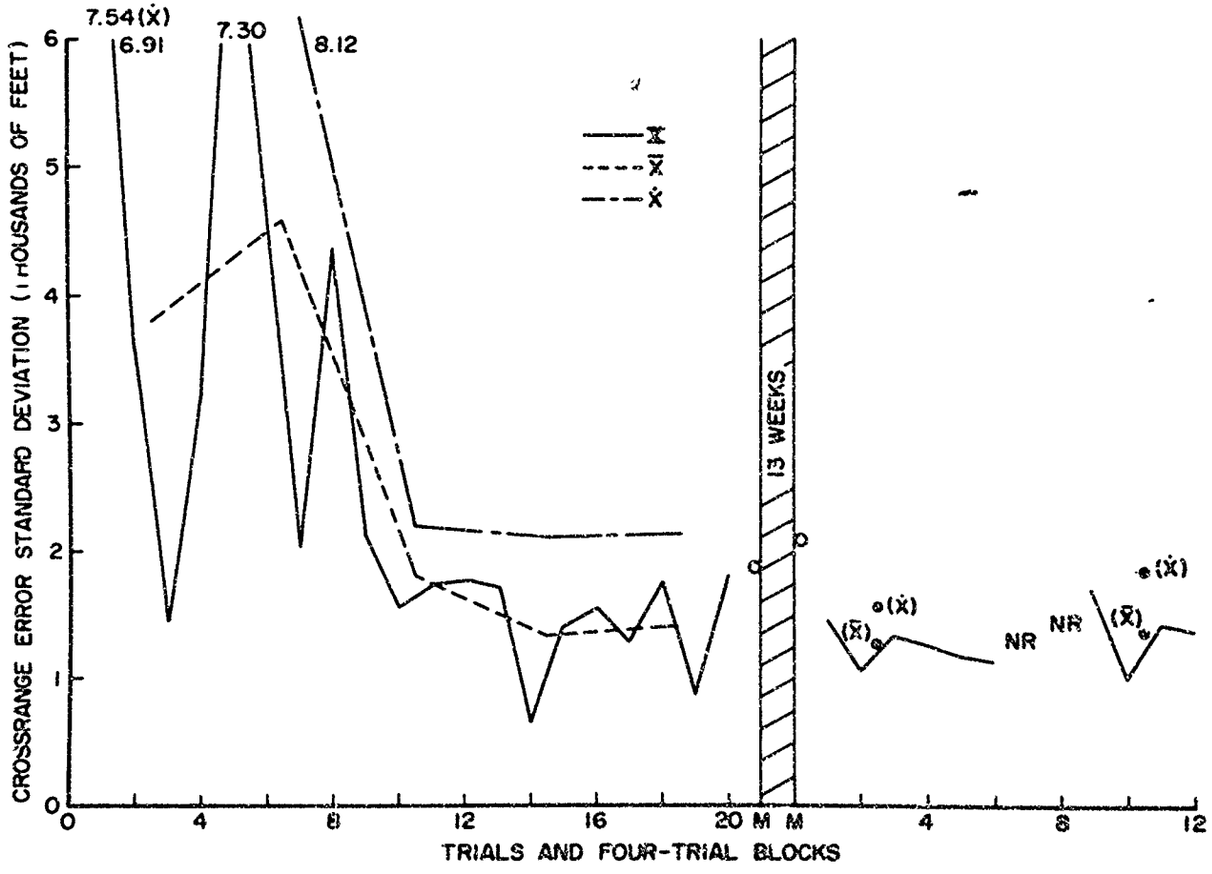
EARTH ENTRY: MEAN ALTITUDE ERROR



EARTH ENTRY: ALTITUDE ERROR STANDARD DEVIATION



EARTH ENTRY: MEAN CROSSRANGE ERROR



EARTH ENTRY: CROSSRANGE ERROR STANDARD DEVIATION

REFERENCES

Cotterman, T.E., A New Approach to Behavioral Measurement and Analysis, AMRL-TR-67-57, Aerospace Medical Research Laboratories, Aerospace Medical Division, Wright-Patterson Air Force Base, Ohio (in press).

Grodsky, M.A., J.A. Mandour, J.T. Warfield, and T.M. Flaherty, Research on Pilot Skill Retention for Manned Space Flight, Research Memorandum 180, Martin-Marietta Corp., Baltimore, Md., August 1964.

Grodsky, M.A., D. Roberts, and J. Mandour, Test of Pilot Retention of Simulated Lunar Mission Skills, Engineering Report 14139, Martin-Marietta Corp., Baltimore, Md., March 1966(a).

Grodsky, M.A., J.A. Mandour, D.L. Roberts, and D.P. Woodward, Crew Performance Studies for Manned Space Flight, Engineering Report 14141, Martin-Marietta Corp., Baltimore, Md., June 1966(b).

Naylor, J.C., and G.E. Briggs, Long-Term Retention of Learned Skills: A Review of the Literature, ASD TR 61-390, Aeronautical Systems Division, Wright-Patterson Air Force Base, Ohio, August 1961.

Peters, C.C., and W.R. van Voorhis, Statistical Procedures and Their Mathematical Bases, McGraw-Hill Book Company, New York, 1940.

Shapero, A., J.I. Cooper, M. Rappaport, K.H. Schaeffer, and C. Bates, Human Engineering and Malfunction Data Collection in Weapon System Test Programs, WADC TR 60-36, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio, February 1960.

Thomas, J.W., Correlation of Operational Reliability with Inherent Reliability of Airborne TACAN Equipment, ASD TDR 62-839, Aeronautical Systems Division, Wright-Patterson Air Force Base, Ohio, November 1962.

Trumbo, D., M. Noble, K. Cross, and L. Ulrich, "Task Predictability in the Organization, Acquisition, and Retention of Tracking Skill," Journal of Experimental Psychology, Vol 70, pp 252-263, 1965.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author) Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio 45433	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b GROUP N/A

3 REPORT TITLE

RETENTION OF SIMULATED LUNAR LANDING MISSION SKILLS: A TEST OF PILOT RELIABILITY

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5 AUTHOR(S) (Last name, first name, initial)

Cotterman, Theodore E., PhD
Wood, Milton E.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April 1967	154	9

8a CONTRACT OR GRANT NO b PROJECT NO. 1710 c Task No. 171003 d	9a. ORIGINATOR'S REPORT NUMBER(S) AMRL-TR-66-222 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10 AVAILABILITY/LIMITATION NOTICES

This document is subject to special export controls and each transmittal to foreign governments or foreign nationals may be made only with prior approval of 6570 Aerospace Medical Research Laboratories (MRHTM), Wright-Patterson AFB, Ohio.

11 SUPPLEMENTARY NOTES Test phases described were conducted by Martin-Marietta Corporation under NASA Contracts NASw-1034 & NASw-1319	12. SPONSORING MILITARY ACTIVITY Aerospace Medical Research Laboratories Aerospace Medical Div., Air Force Systems Command, Wright-Patterson AFB, Ohio

13 ABSTRACT Four crews of three pilots were tested on a simulated 7-day lunar landing mission at intervals of 4, 8, 9, and 13 weeks, after original training. The 6 weeks of training culminated in real time performance of the mission but, for the skill retention test the mission was compressed into a single 13-hour workday by omitting less significant tasks and waiting periods. Following the test 1 or 3 days of additional training on selected mission phases was given all crews. The analysis of results focused attention on individual and crew performance at the end of training, in the skill retention test mission, and in the following retraining trials, as represented by 22 selected flight control parameters distributed over nine mission phases. Using novel analytic techniques the levels of performance observed were given as reliabilities, or probabilities of success in meeting hypothetical criteria for the parameters. Also, for greater sensitivity to changes in capability, test and retraining performances were alternatively given as probabilities of success in meeting the level of performance estimated achievable by each individual in 95% of his performances at the end of training. The obtained probabilities are taken to indicate (1) lack of direct practice of critical tasks over 8 weeks or more in space missions will result in unacceptable skill deterioration unless design and operational planning are remedied, (2) aerospace research pilots are capable of performing the type of mission used in this study, if extreme care is given to their training and individual performance reliability is demonstrated. Further research on skill retention is indicated and advantages of the analytic methodology used are stated.

DD FORM 1 JAN 64 1473

Security Classification

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Training						
Skill retention						
Pilot reliability						
Behavioral measurement						
Lunar landing mission						
Manned orbital laboratory (MOL)						

INSTRUCTIONS

1. ORIGINATING ACTIVITY Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13 ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14 KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Security Classification

SUPPLEMENTARY

INFORMATION

IDENTIFICATION	FORMER STATEMENT	NEW STATEMENT	AUTHORITY
AD-817 232 Aerospace Medical Research Labs., Wright-Patterson AFB, Ohio. Rept. no. AMRL-TR-66-222 Apr 67	No Foreign without approval of Aerospace Medical Research Labs., Attn: MRHTM, Wright-Patterson AFB, Ohio.	No Foreign without approval of Human Resources Laboratory, Attn: HRT, Wright-Patterson AFB, Ohio.	AMRL, USAF ltr, 22 Dec 69